

Discussion Session-I

CS771: Introduction to Machine Learning

Purushottam Kar



Miscellaneous Queries

Sept 3, 2017



Miscellaneous Queries

- *Can you discuss the overall structure of ML? Where does Gradient Descent fit into this?*

Look at lecture 1 for an overview. Gradient descent is a solution strategy for optimization problems as those arise in MAP, MLE and FA approaches.

- *There are so many different formulae and equations. How can we remember everything?*

You do not have to remember everything! Examinations (and life in ML) is open notes! However, you do have to appreciate the utility of these eqns.

- *What kind of questions will be asked in mid sem?*

Look at the assignments, questions in reference material for examples.

Miscellaneous Queries

- *Any references for matrix calculus?*

Search for “The Matrix Cookbook” online.

- *Any references for multivariate calculus?*

<http://home.iitk.ac.in/~psraj/mth101/>

<http://home.iitk.ac.in/~shivampa/website/acads/mth101.html>

<http://home.iitk.ac.in/~rishjha/mth101.html>

Any others you come across on MIT OCW, Coursera etc.

- *Any references for linear algebra?*

<http://home.iitk.ac.in/~peeyush/102A/MTH-102A.html>

<http://home.iitk.ac.in/~aralal/mth102a.htm>

<http://home.iitk.ac.in/~rishjha/mth102.html>

<http://home.iitk.ac.in/~shivampa/website/acads/mth102.html>

Any others you come across on MIT OCW, Coursera etc.

Miscellaneous Queries

- *Any references for probability theory basics (distributions etc)?*

<http://home.iitk.ac.in/~neeraj/mso201a/mso201a.htm>

<http://nptel.ac.in/courses/111105041/>

Any others you come across on MIT OCW, Coursera etc.

- *How is it possible to optimize F-measure although it is a complicated expression with the harmonic mean? Can you explain it in layman terms?*

Sort of, it involves taking a complicate problem like F-measure optimization and casting it into a sequence of simpler problems like logistic regression. If you wish to know about it, come and talk to me.

- *What does the notation $\prod_{i=1}^n v_i$ mean?*

It means the product of the real values v_1, v_2, \dots, v_n .

Miscellaneous Queries

- *How to obtain closed form solution from the big summation that we always have?*

Use the first order optimality condition – set the gradient to zero!

- *I don't see why the solution to ridge regression is $\hat{\mathbf{w}} = (XX^T + \lambda \cdot I)^{-1}X\mathbf{y}$. Can't we get a more intuitive form to do away with the math?*

Deep and visual intuition gets progressively harder to achieve in higher dimensions. Visualization of *effects* of this form admirable e.g. what happens when λ is increased or set to zero etc but after a while, it is wise to just get comfortable with the math (previous slide links). Think of the solution to a quadratic equation in one-dimension.

$$y(x) = a \cdot x^2 + b \cdot x + c \text{ has roots } x_0 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Assignment 1

Sept 3, 2017



Assignment 1 related queries

- Please check Piazza for announcements regarding Assignment 1
- Use of Scikit-learn, SciPy, Shogun routines (except LMNN) not allowed in Assignment 1 Problem 1.6

- *What is the difference between L_2 and feature regularization? Aren't they the same mathematically?*

Error in problem sheet $\sum_{j=1}^d \alpha_i \mathbf{w}_i^2$

Not really, the L_2 regularizer is $\|\mathbf{w}\|_2^2 = \sum_{i=1}^d \mathbf{w}_i^2$ and feature regularization looks like $\sum_{i=1}^d \alpha_i \mathbf{w}_i^2$, it places different emphasis on different features.

- *For problems 1.4 and 1.5, where is the dataset to work with?*
No dataset, you have to show these results formally (i.e. mathematically).

Probabilistic Machine Learning

Sept 3, 2017



PML Queries

- *What exactly is PML? Why PML? is it better than other non-PML approaches?*

PML is one of the many approaches to machine learning. There are situations where it is preferable and others where non-PML approaches are preferable. Understanding these subtleties takes time and practice.

- *Do all loss functions correspond to some likelihood model?*

In general yes, but it is difficult to find the exact, closed form solution for that likelihood model. It was only very recently that one was found for hinge loss (see Piazza comment by Prof. Piyush Rai).

- *Given a regularized optimization problem, is there a generic way to extract prior from it?*

No. Not if you want a simple closed form expression.

PML Queries

- *What is meant by “design a likelihood distribution”?*

This is likely asked in the context of the assignment. A likelihood distribution is a probability distribution of the form $\mathbb{P}[y \mid \mathbf{x}, \mathbf{w}]$. For example, in binary classification, we used the logistic likelihood model

$$\mathbb{P}[y \mid \mathbf{x}, \mathbf{w}] = \sigma(y\langle \mathbf{w}, \mathbf{x} \rangle), y \in \{-1, +1\}$$

- *In section 9.7 of DAU, how come the probability of prior be taken as Gaussian?*

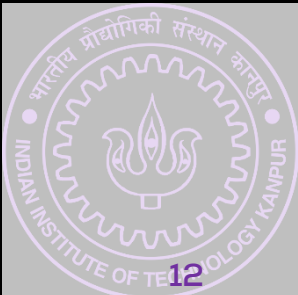
The Gaussian prior is only one of the many possible prior choices.

- *What does the Beta prior mean? Where did this formula come from? What are α, β ?*

The beta distribution is a distribution over the interval $[0,1]$. It is a so-called *parameterized* distribution. There are infinitely many Beta distributions, one for each value of the parameters $\alpha, \beta > 0$. See Piazza post for more.

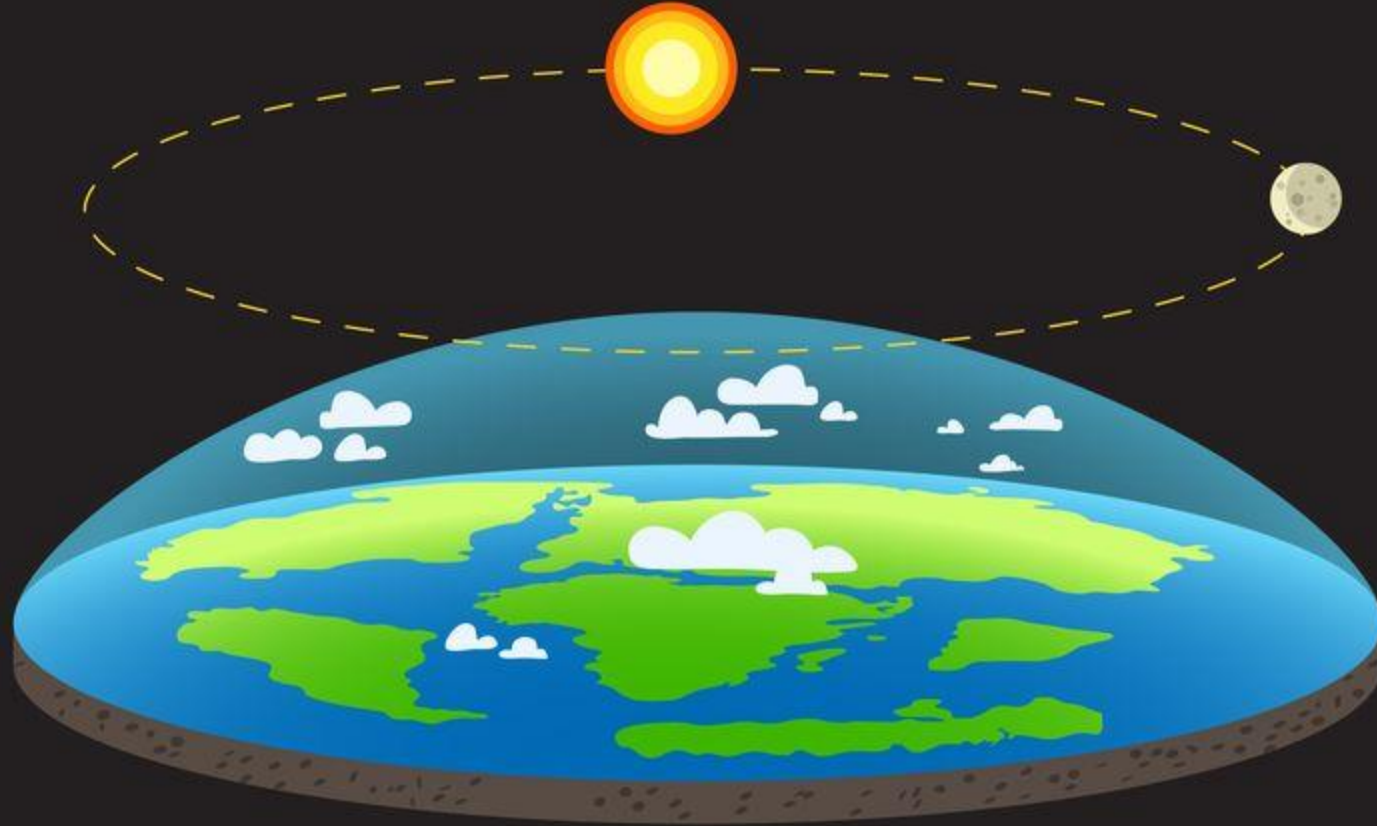
In the beginning ...

Sept 3, 2017

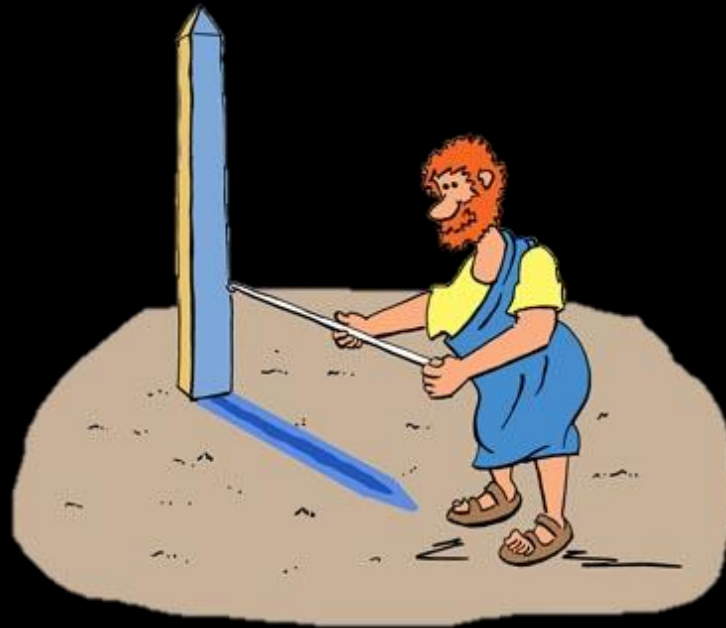


CS771: Intro to ML

Prior belief is acquired ...



Arguments are made ...

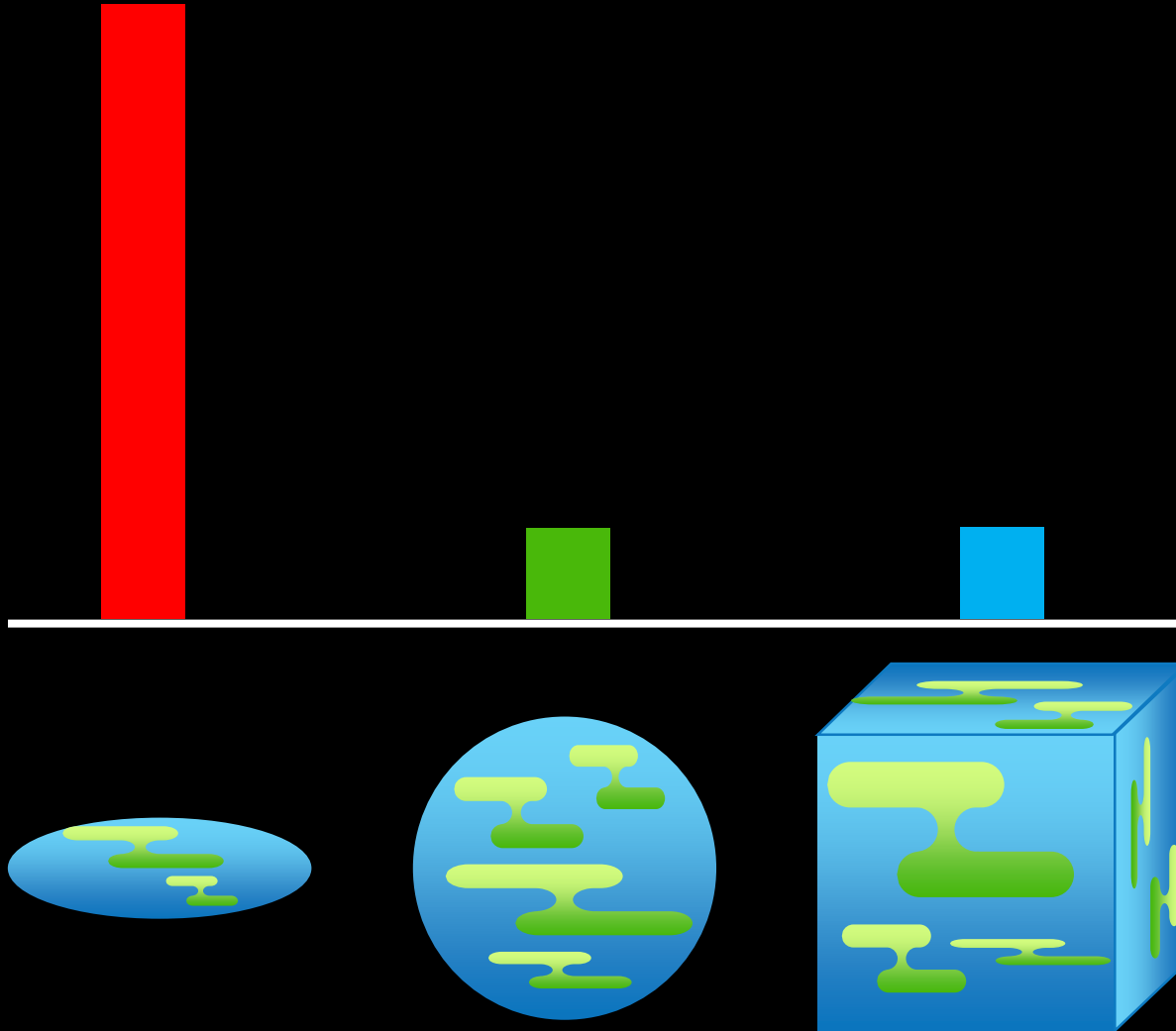


Sept 3, 2017

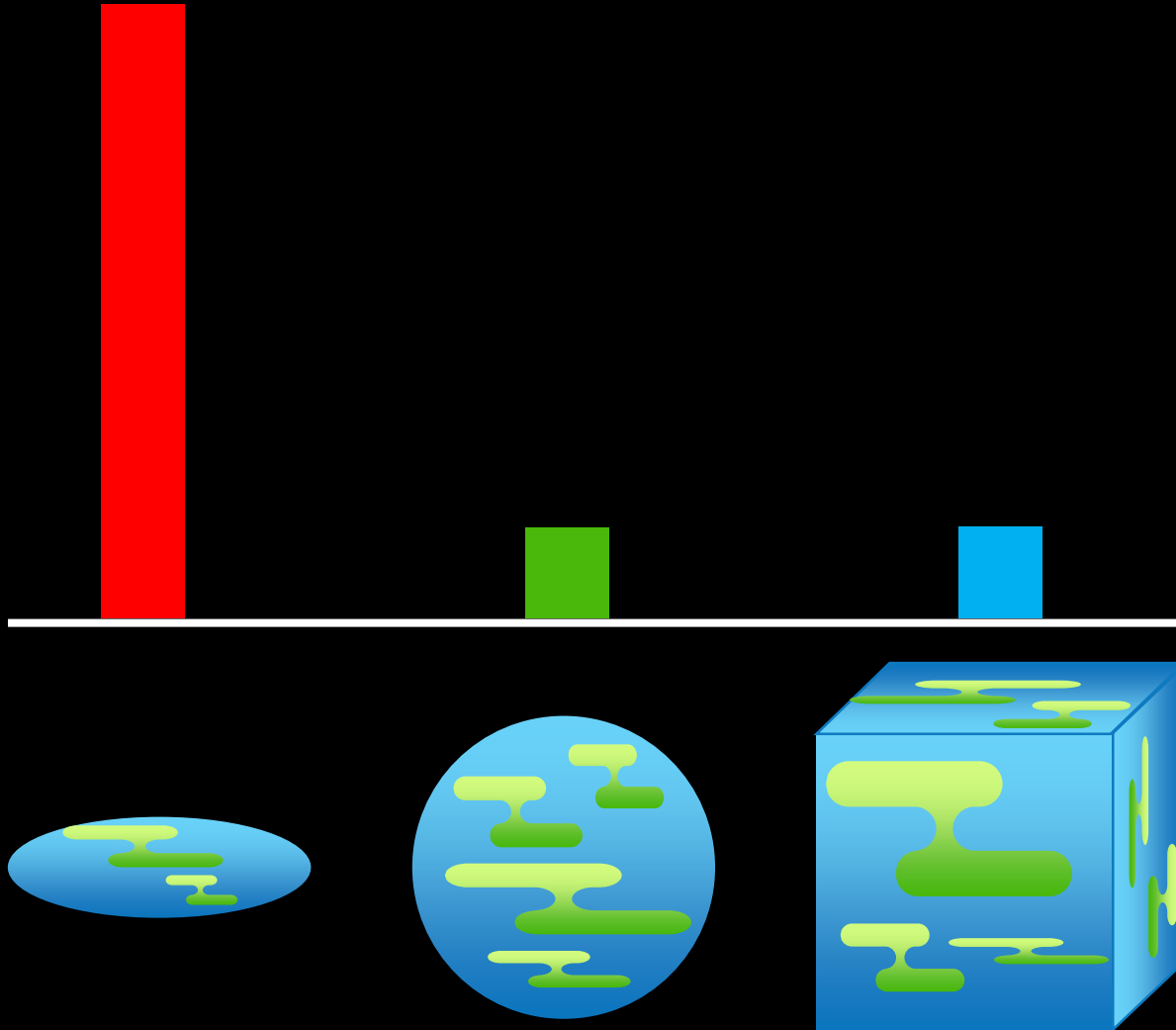
theflatearthsociety.org, coopertoons.com, nasa.gov



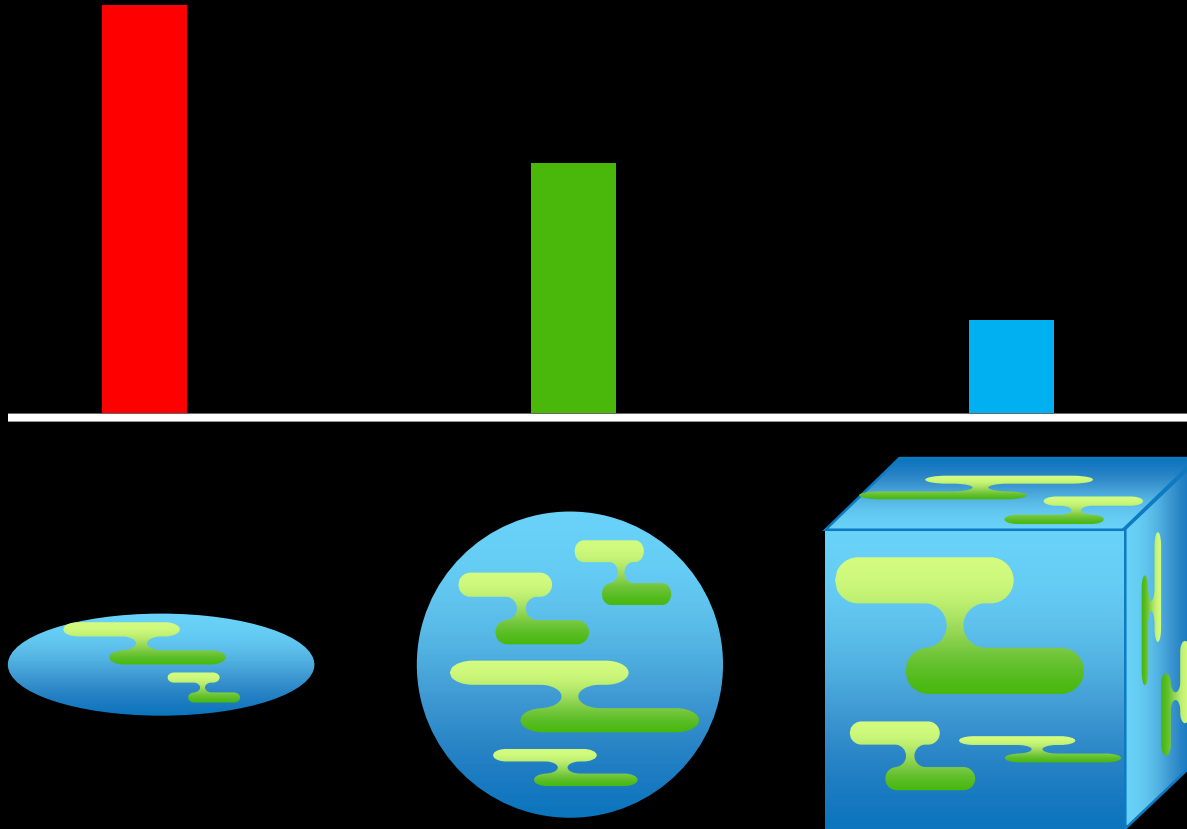
The Belief is altered ...



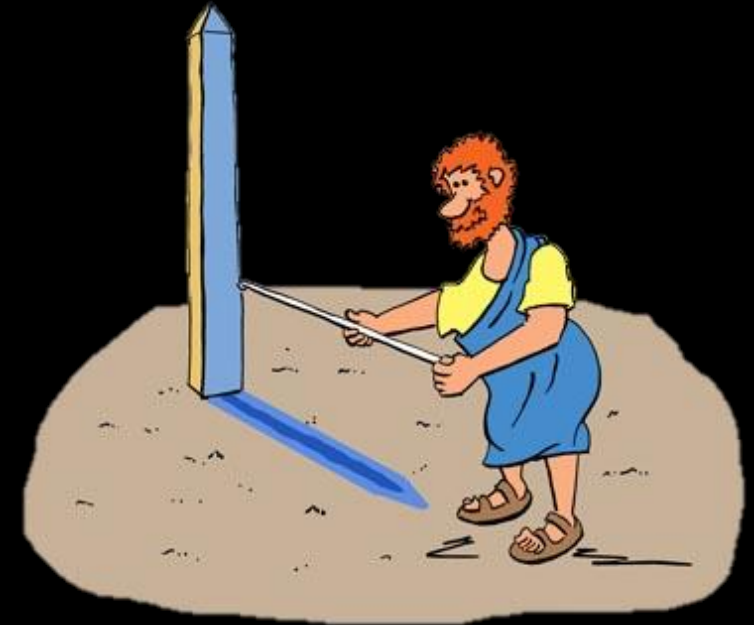
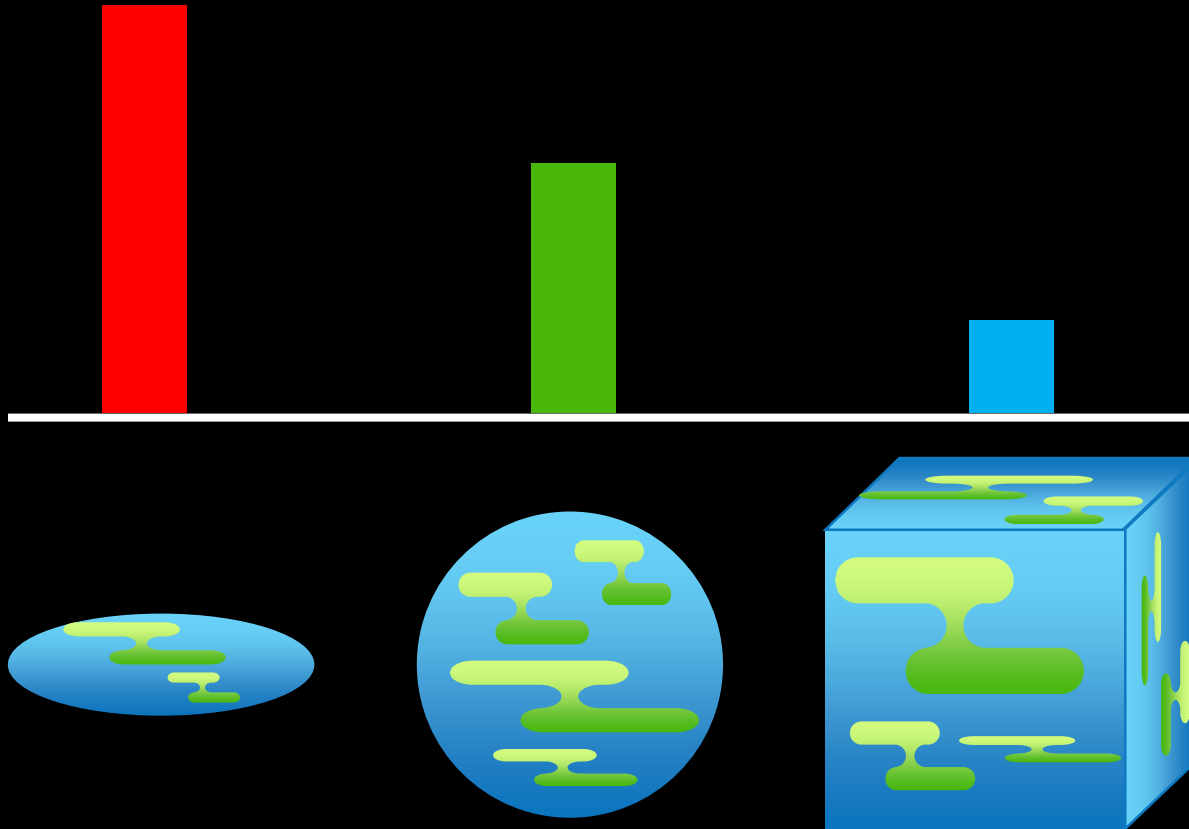
The Belief is altered ...



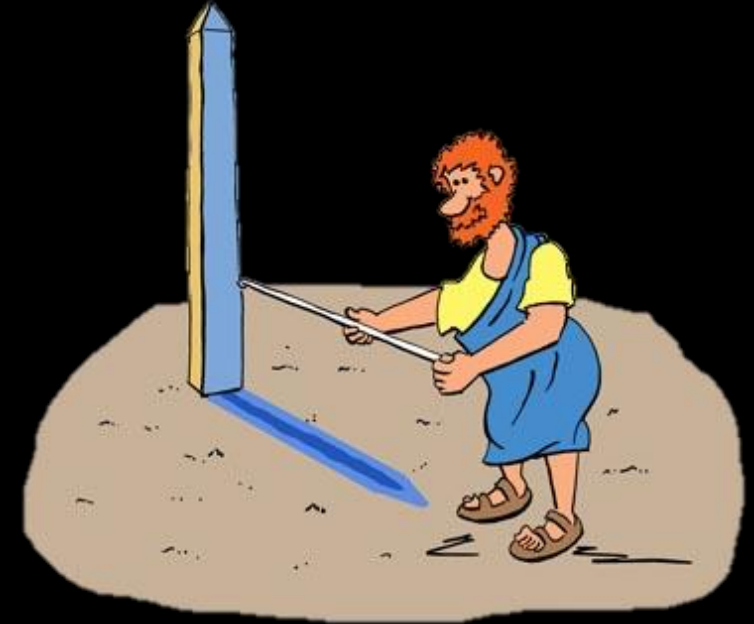
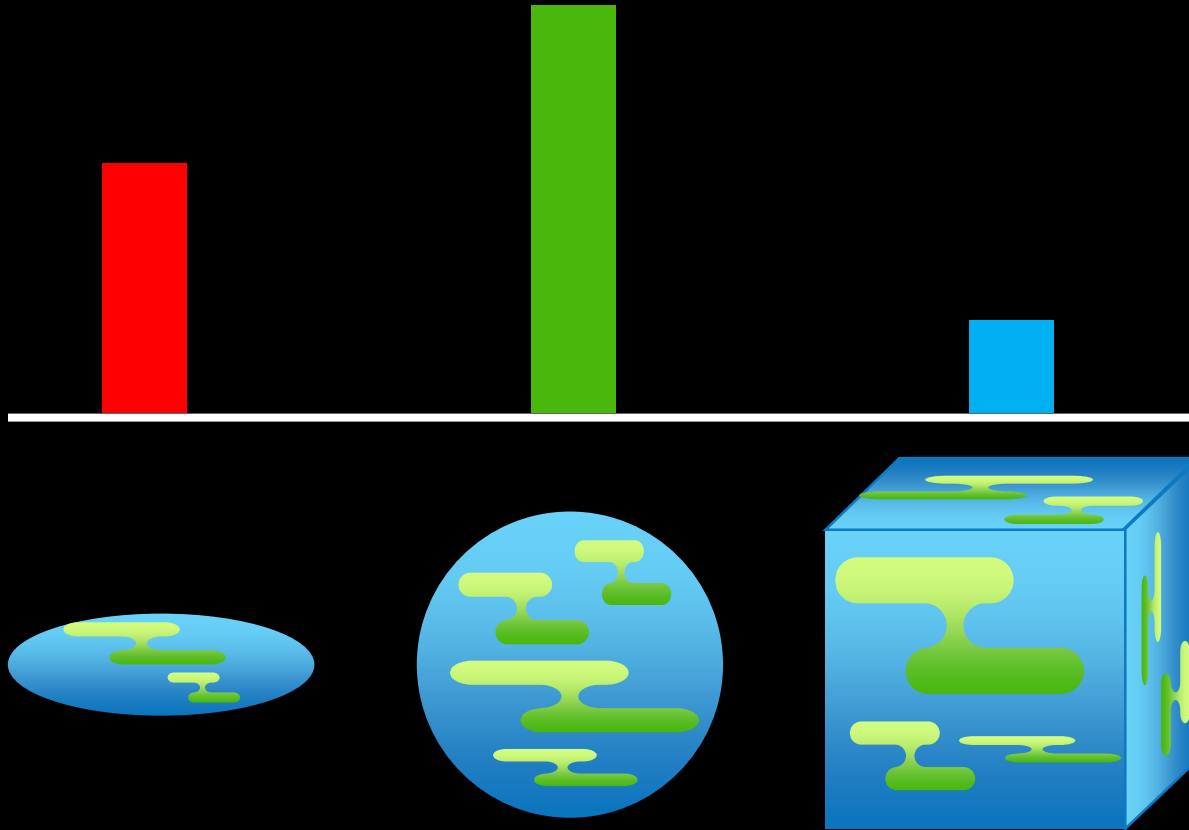
The Belief is altered ...



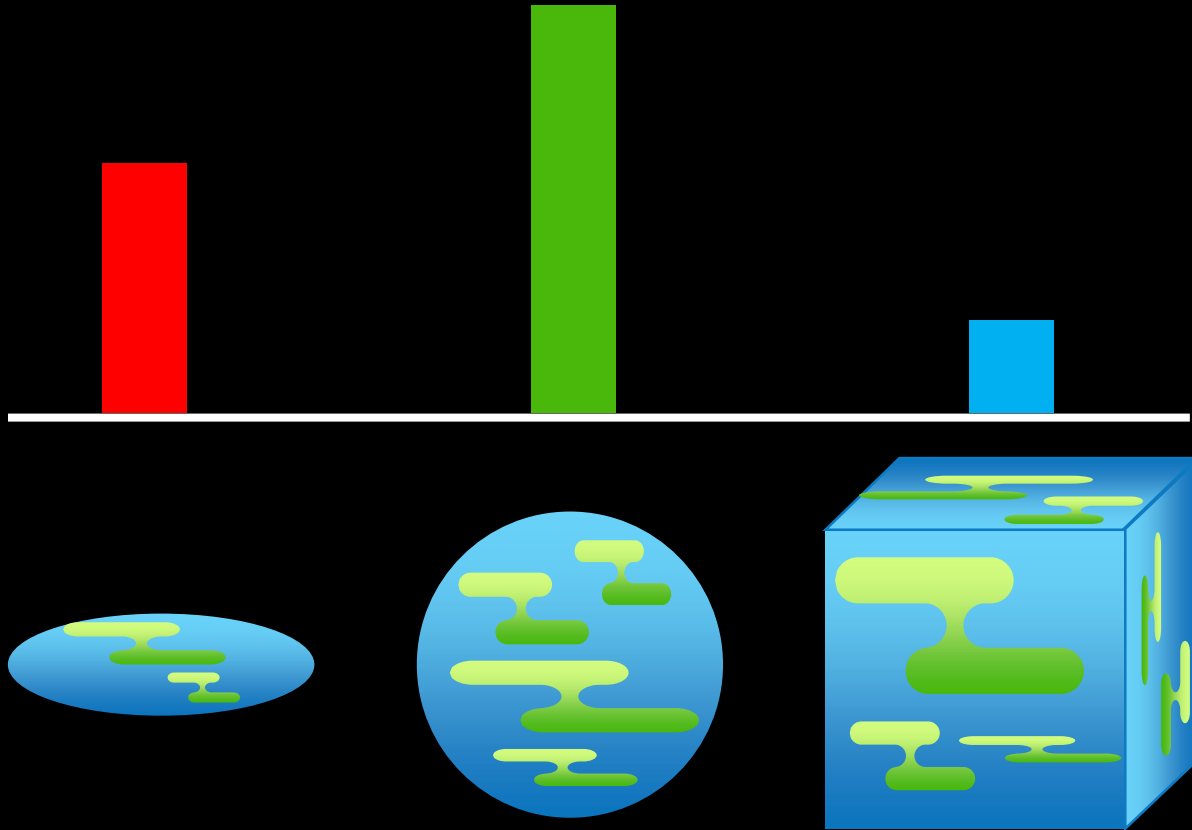
The Belief is altered ...



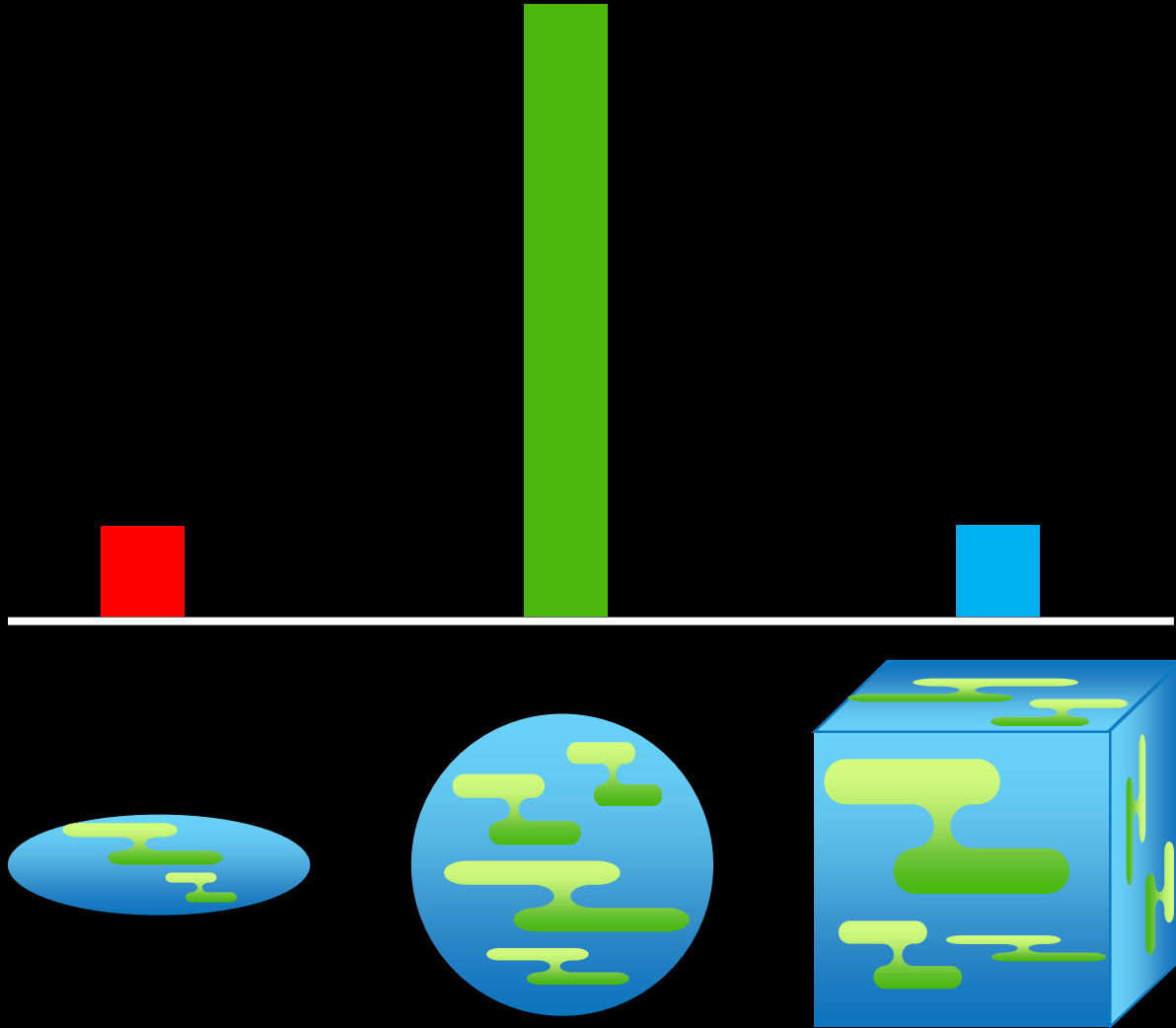
The Belief is altered ...



The Belief is altered ...

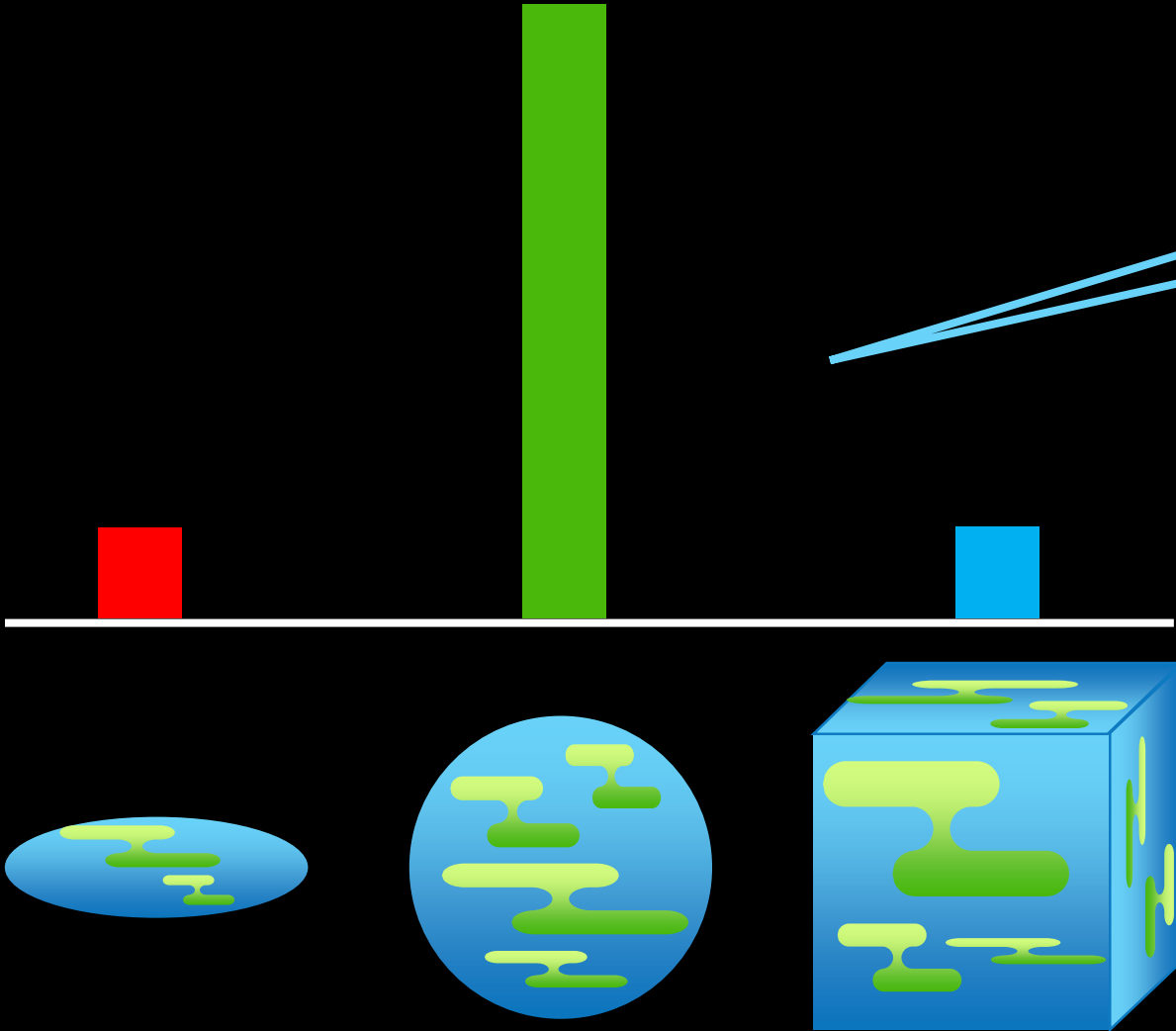


The Belief is altered ...

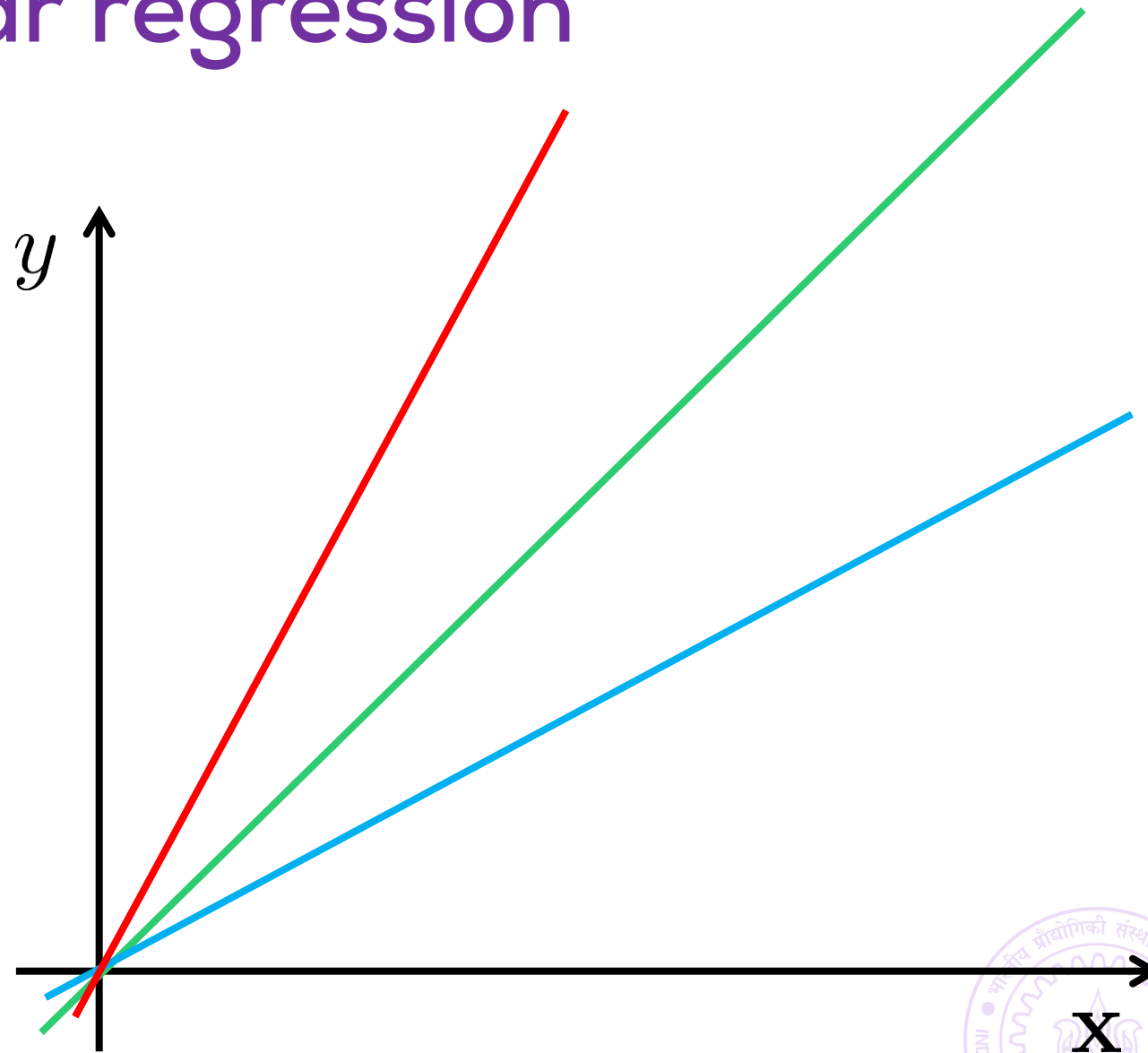


The Belief is altered ...

Probability distributions popular way of representing belief – not the only way though

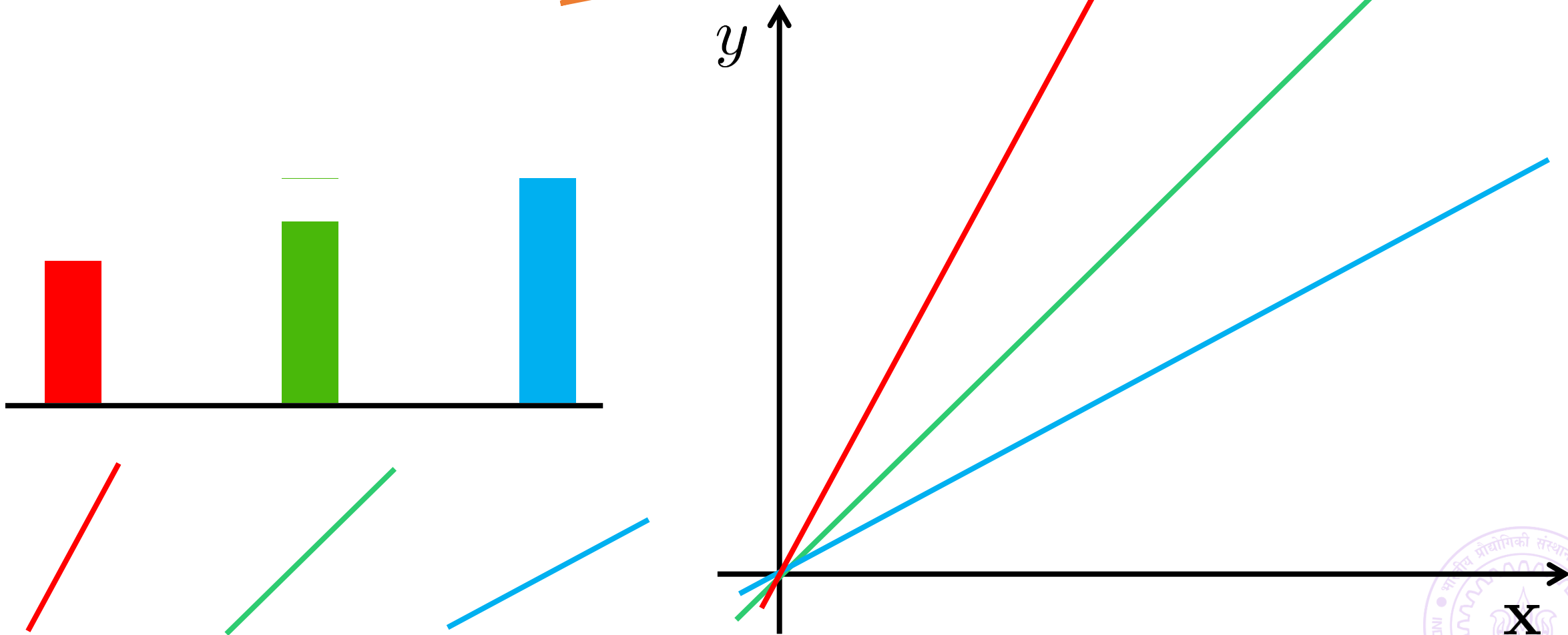


The PML view of linear regression



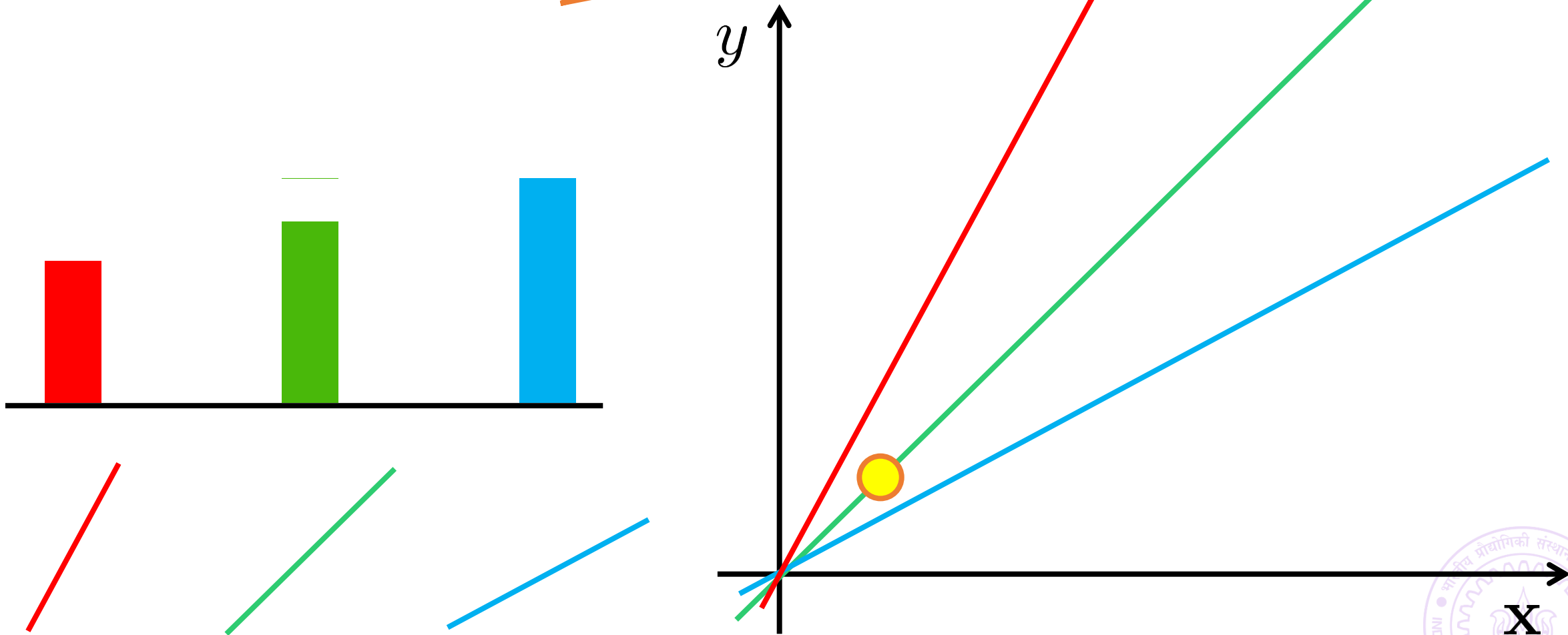
The PML view of linear

Non-uniform
prior distribution



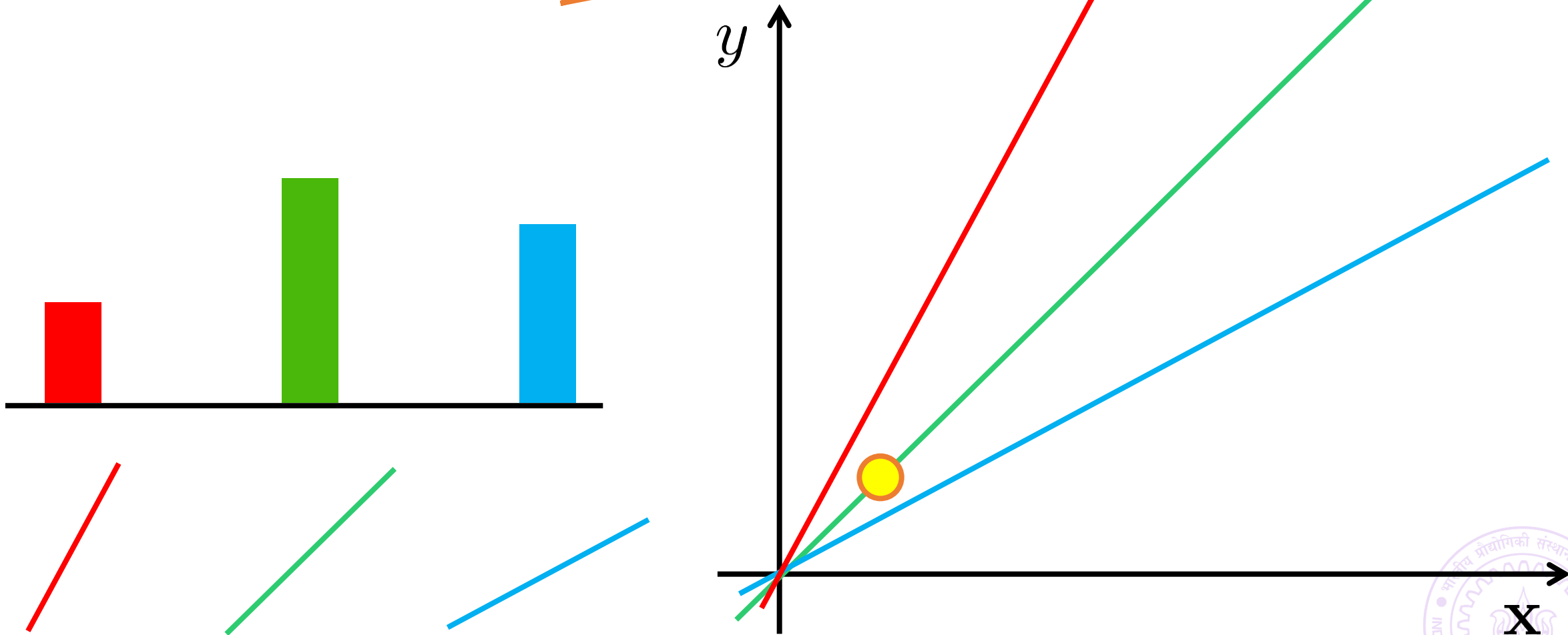
The PML view of linear

Non-uniform
prior distribution



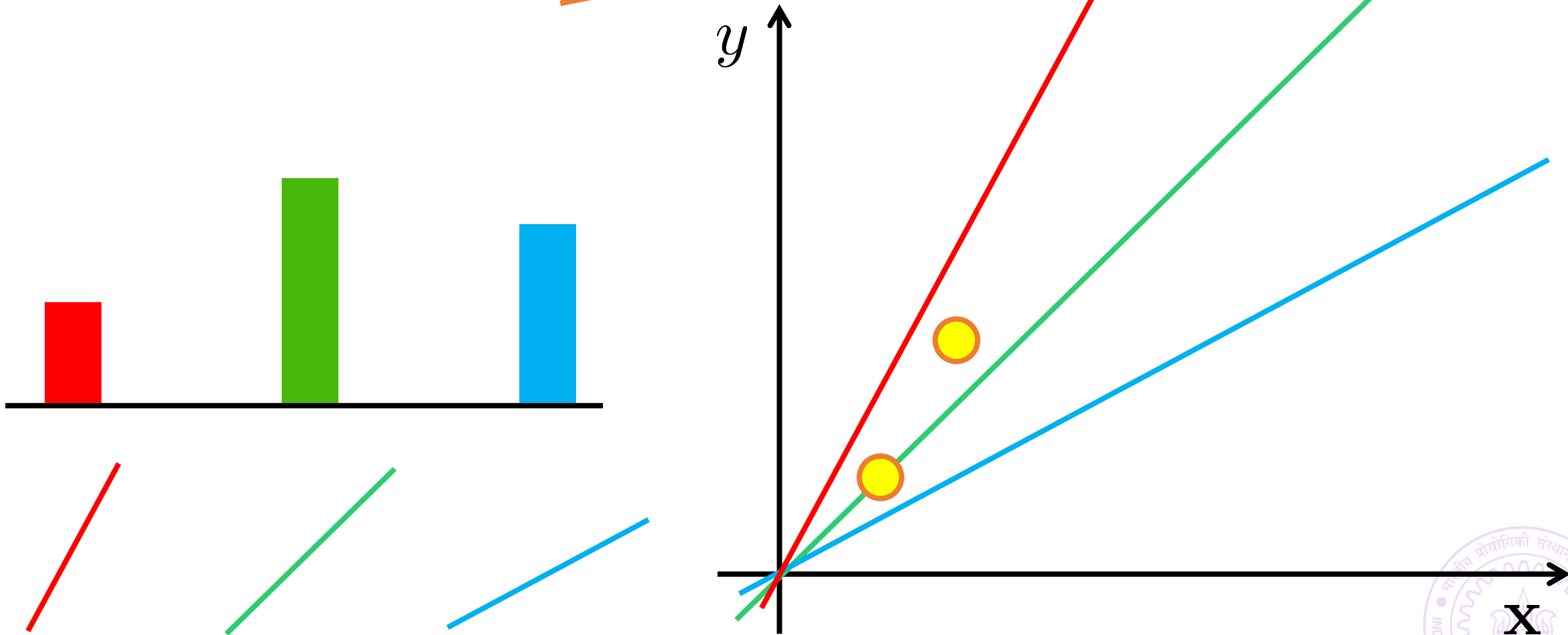
The PML view of linear

Non-uniform
prior distribution



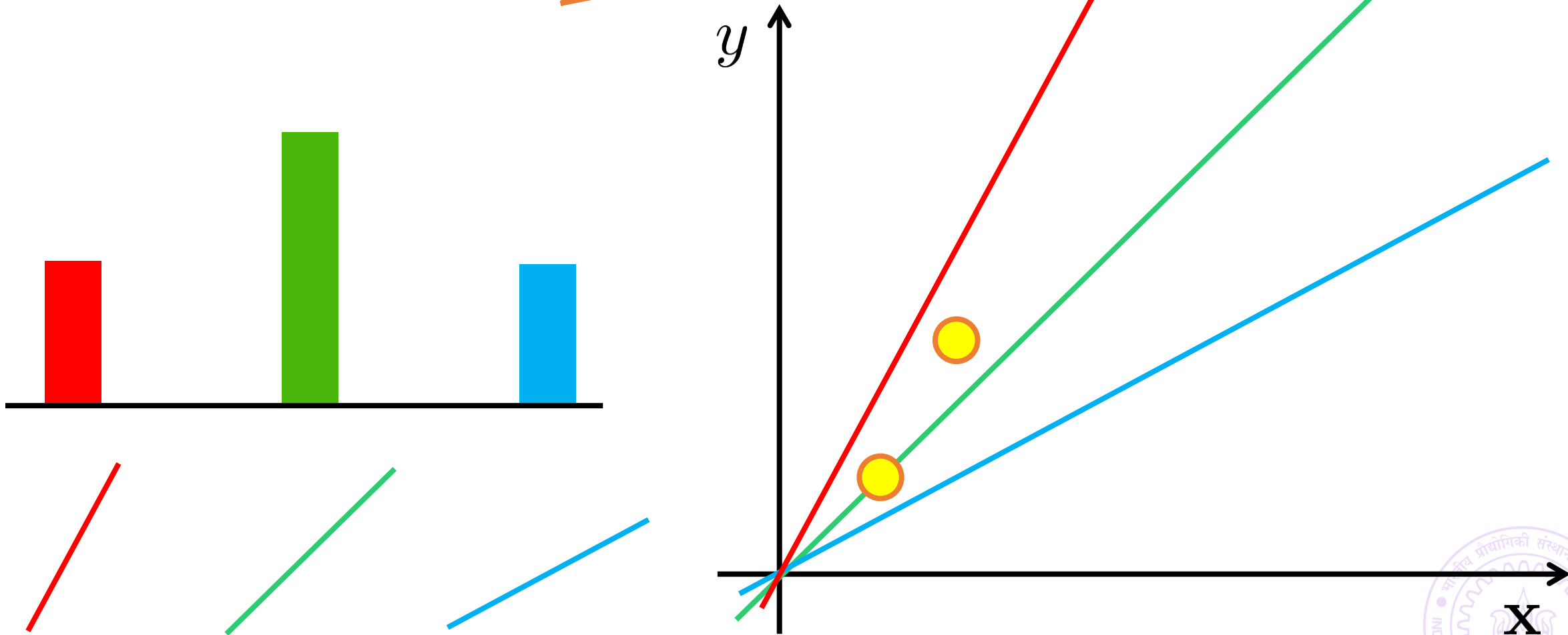
The PML view of linear

Non-uniform
prior distribution



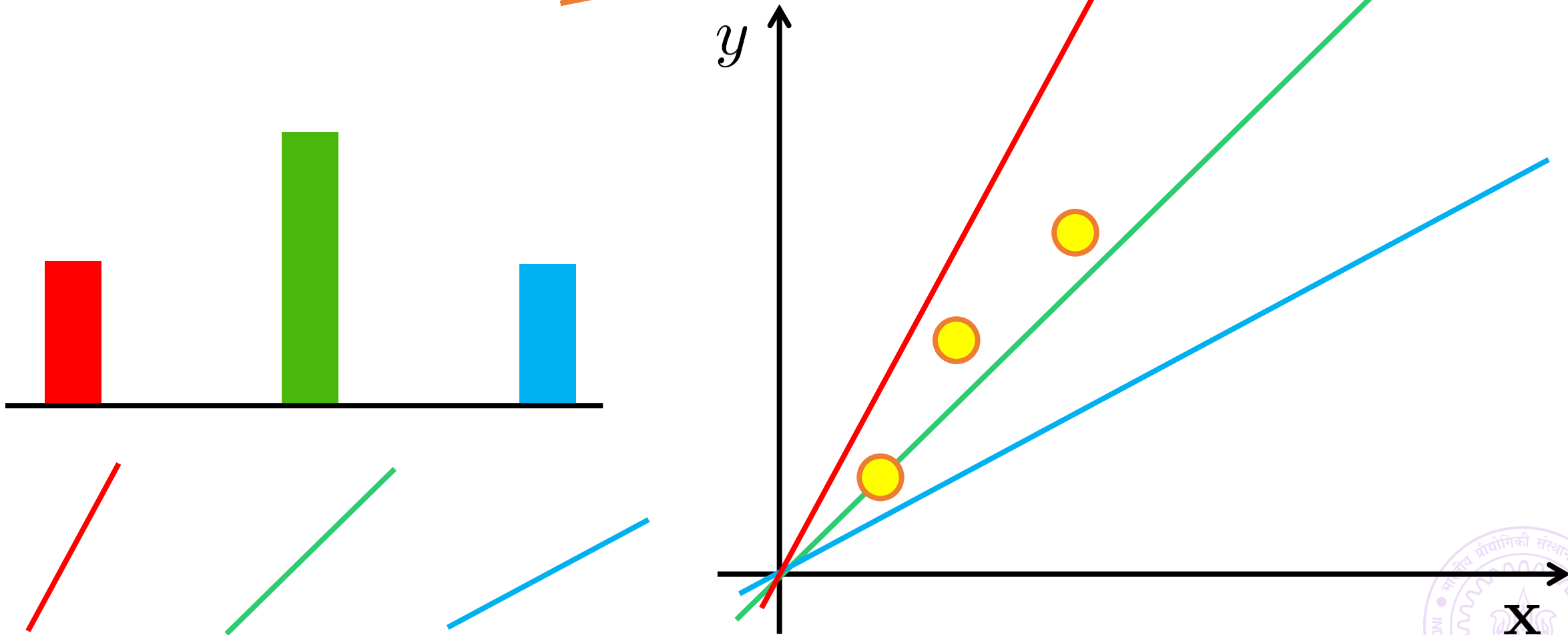
The PML view of linear

Non-uniform
prior distribution



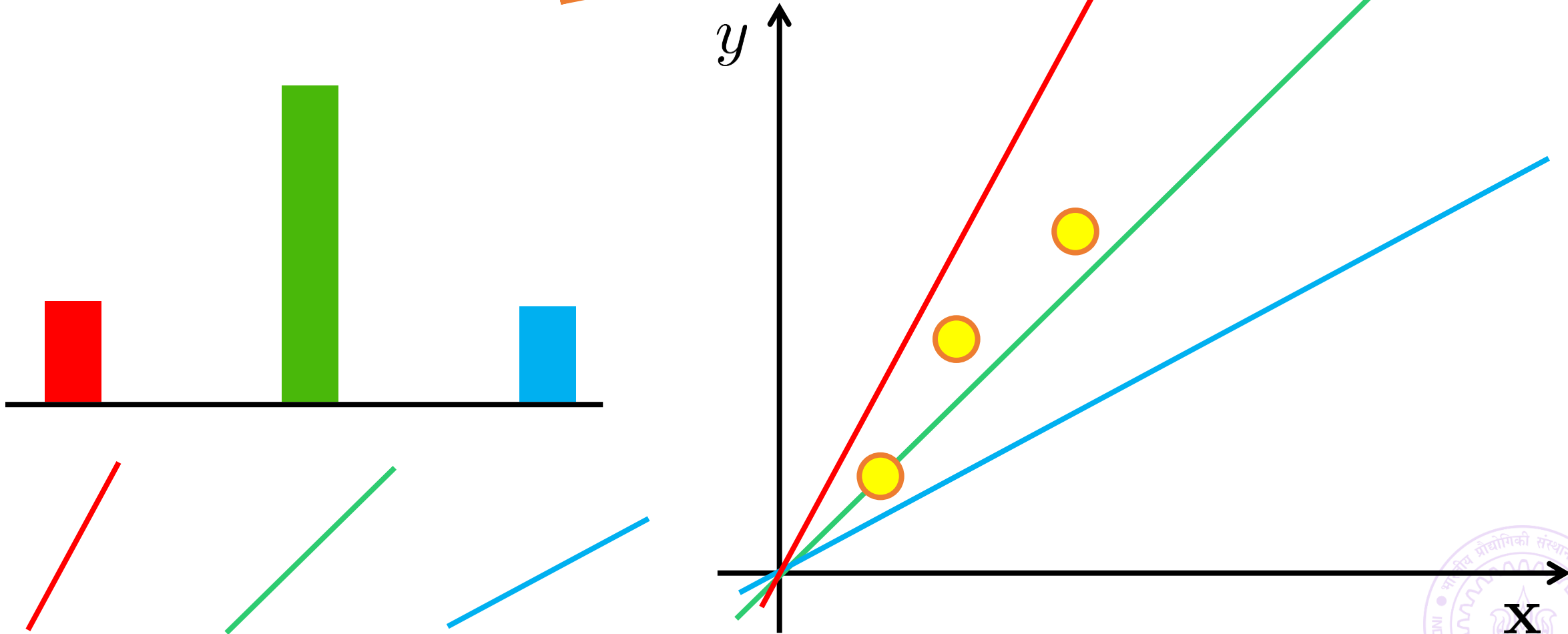
The PML view of linear

Non-uniform
prior distribution



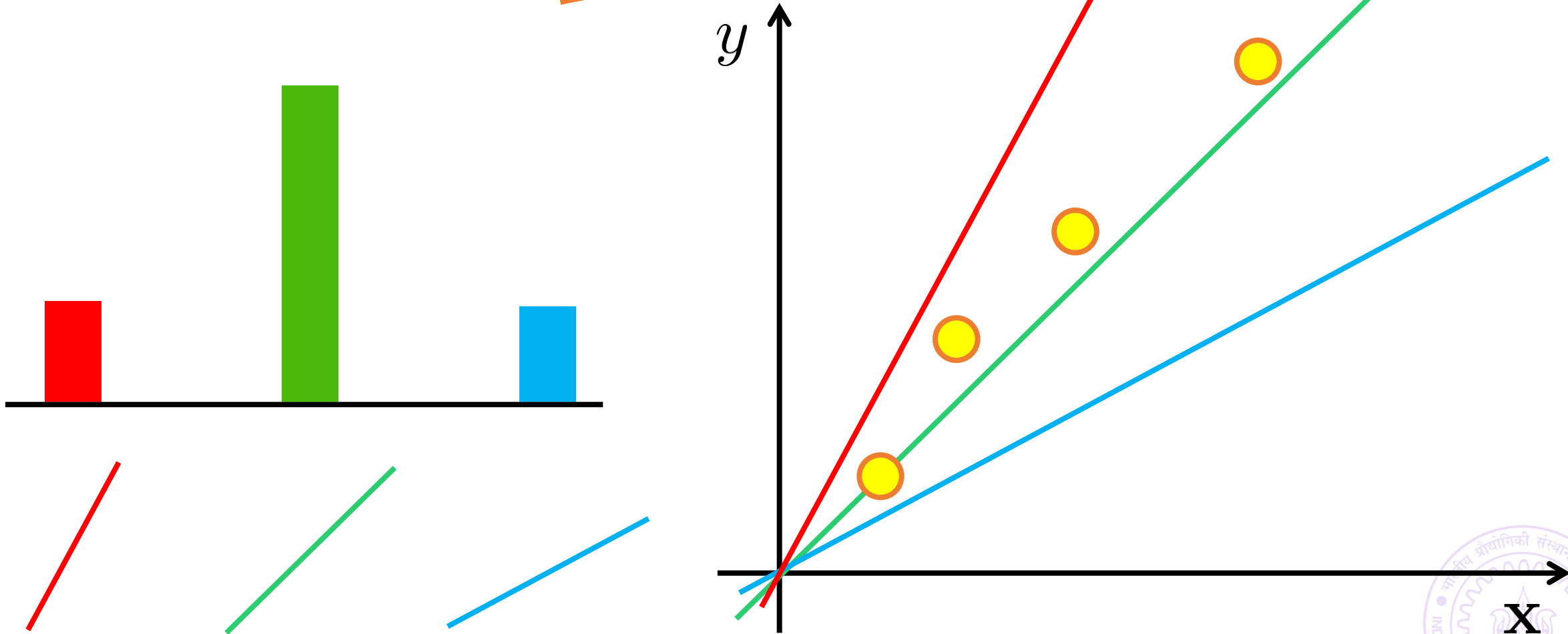
The PML view of linear

Non-uniform
prior distribution



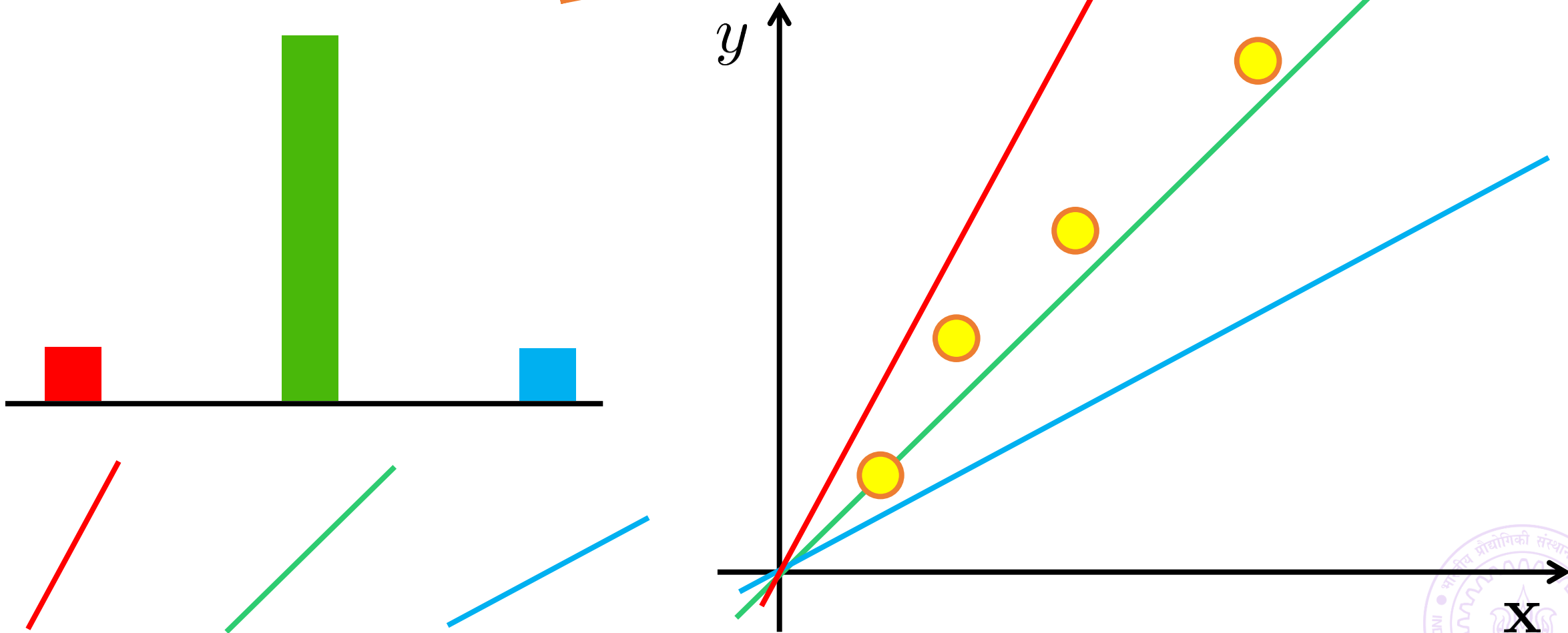
The PML view of linear

Non-uniform
prior distribution



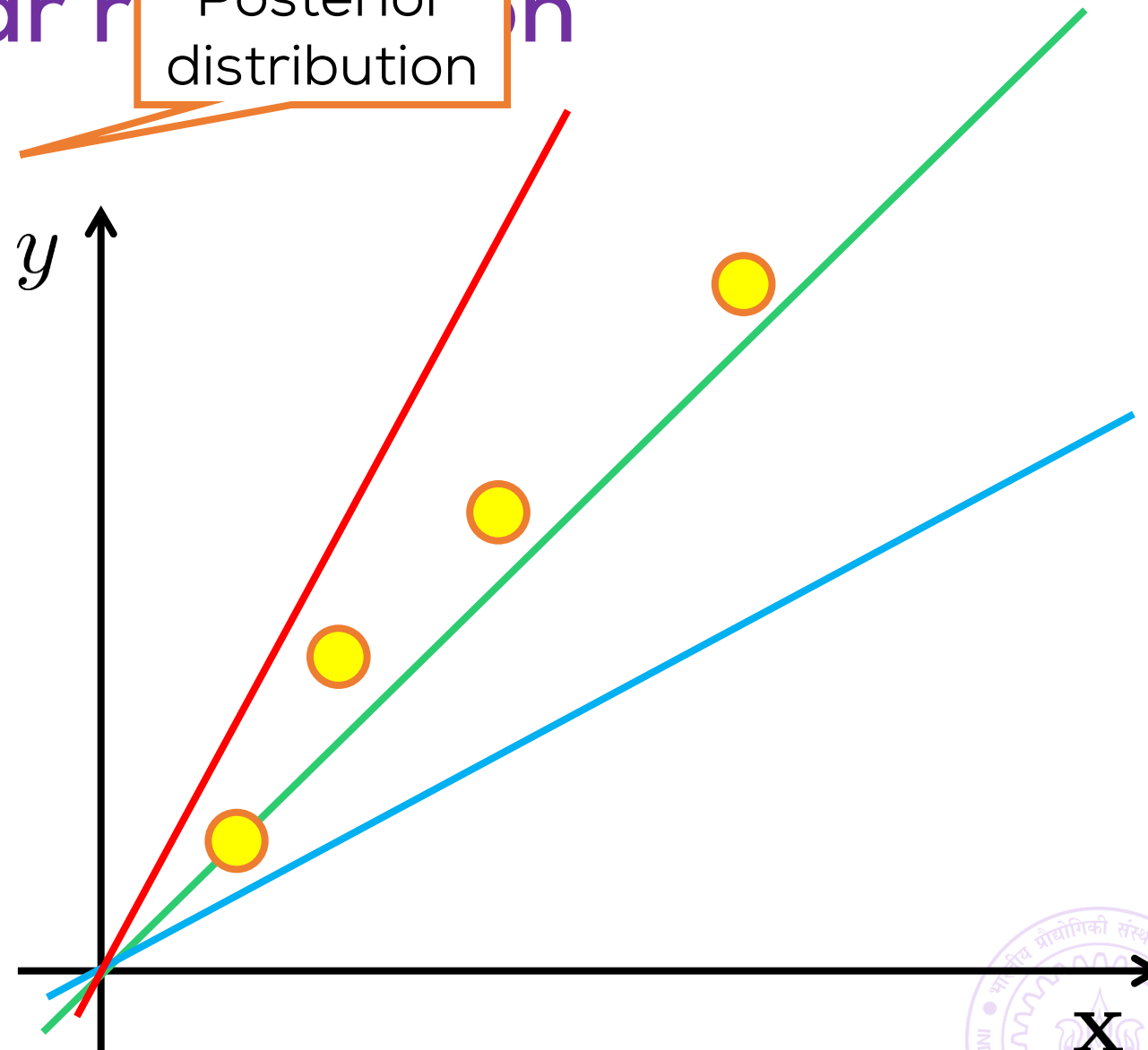
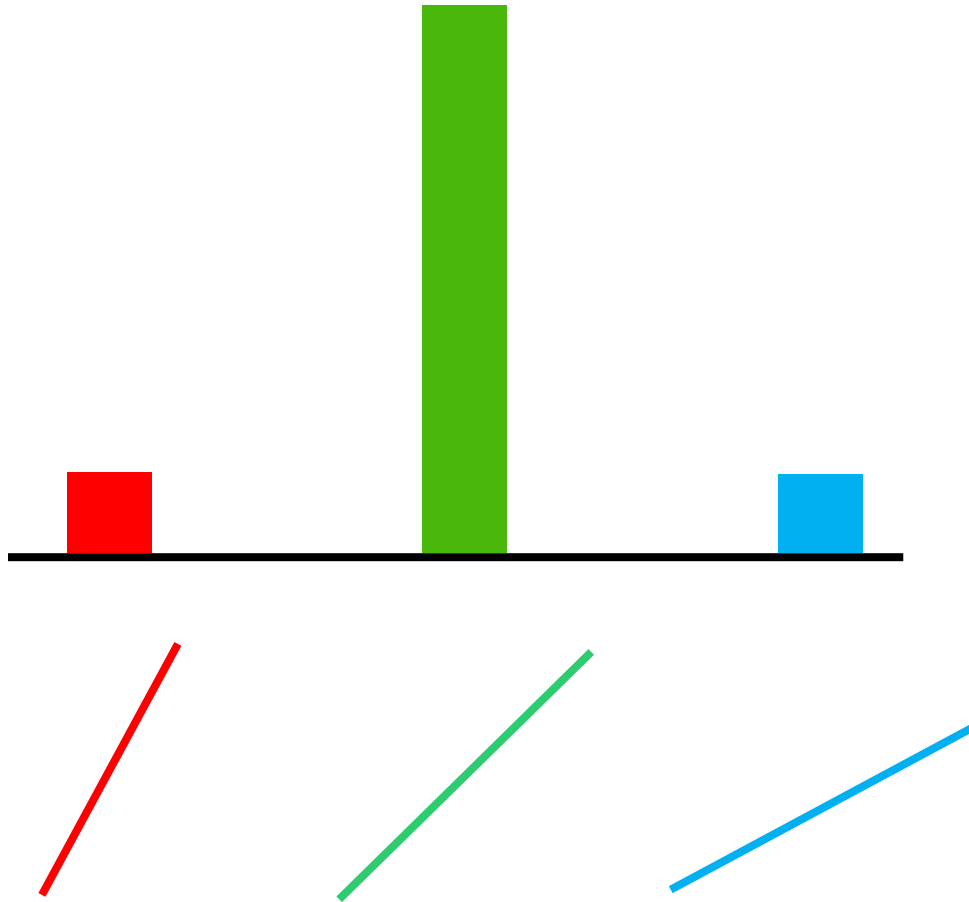
The PML view of linear

Non-uniform
prior distribution

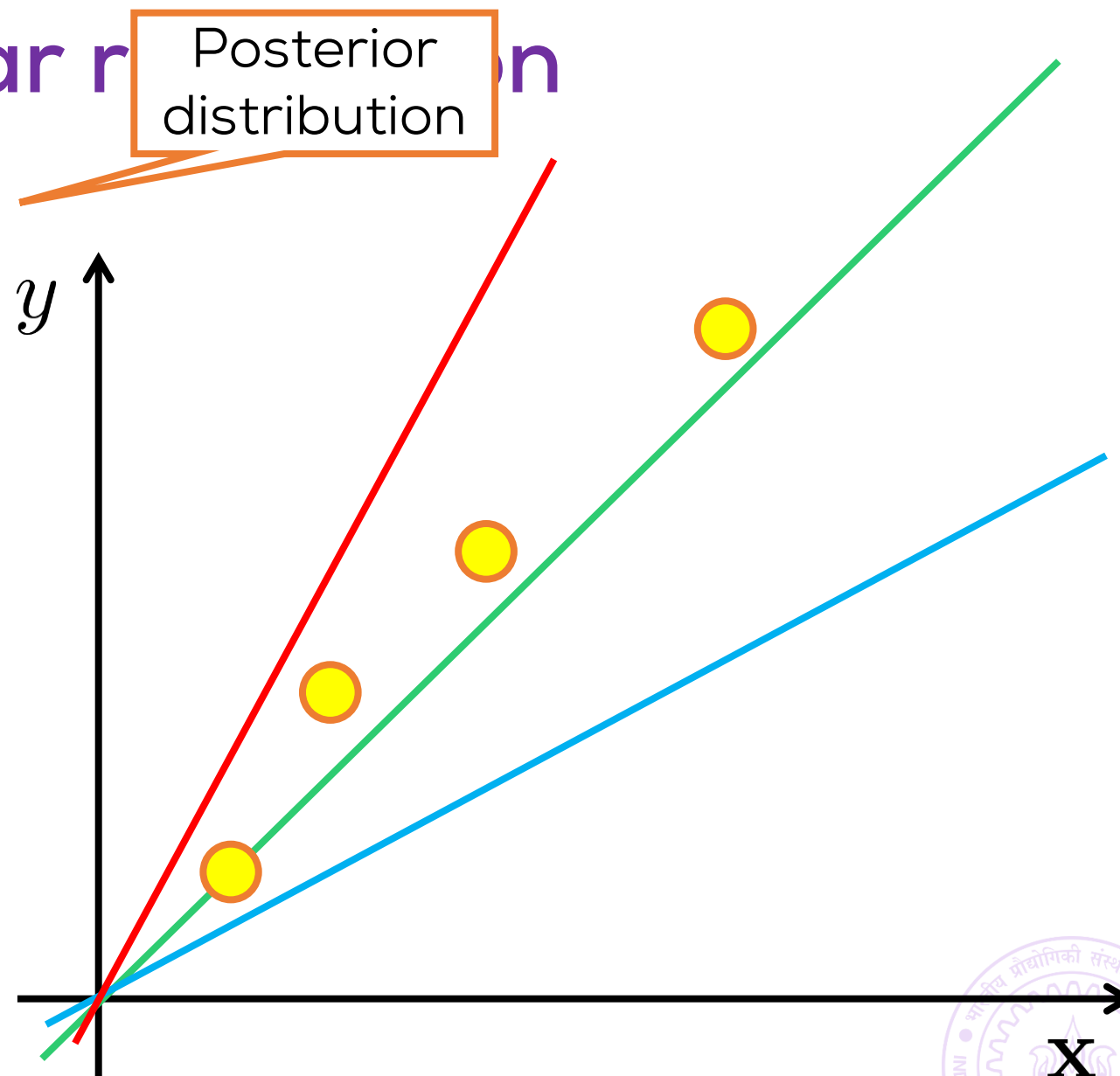
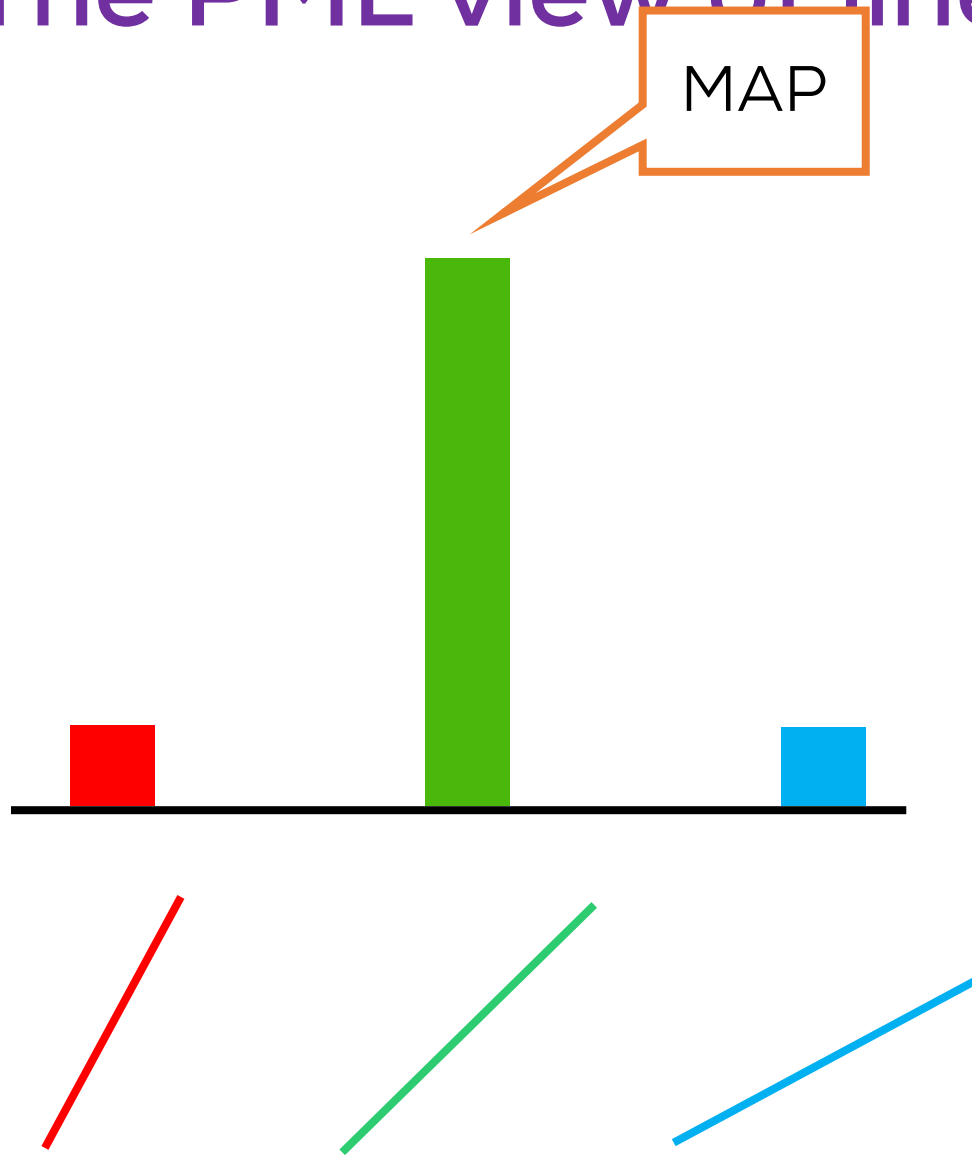


The PML view of linear regression

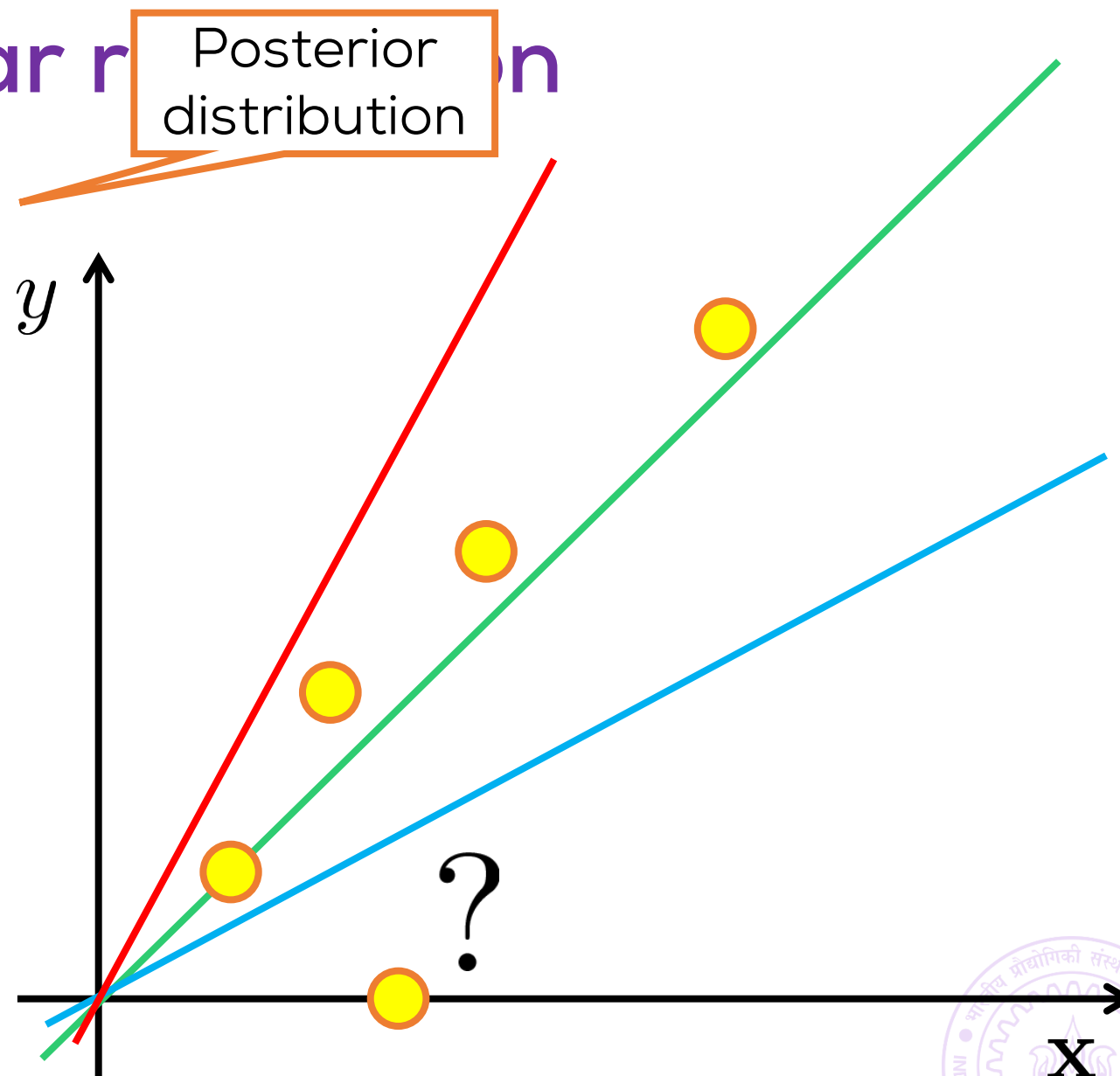
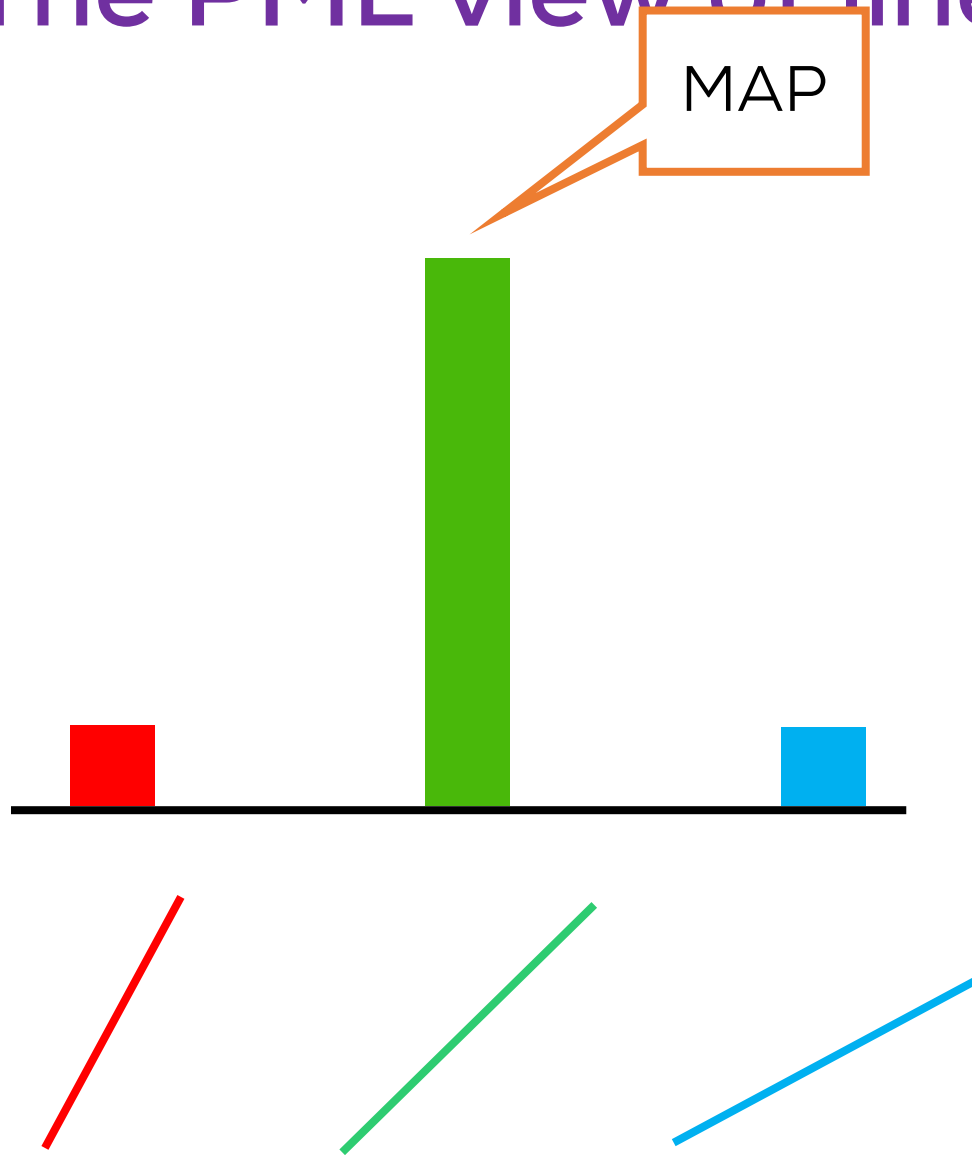
Posterior distribution



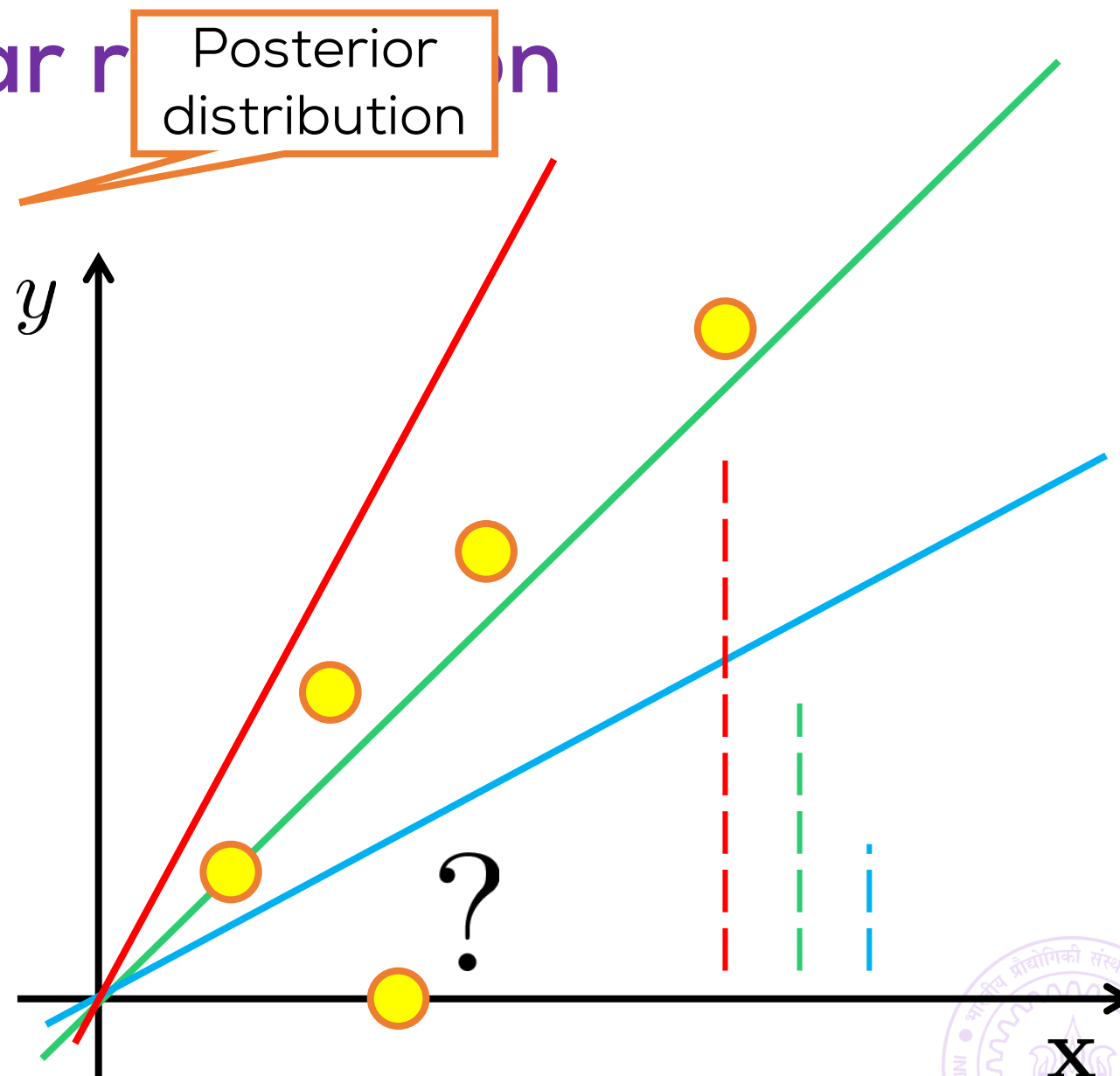
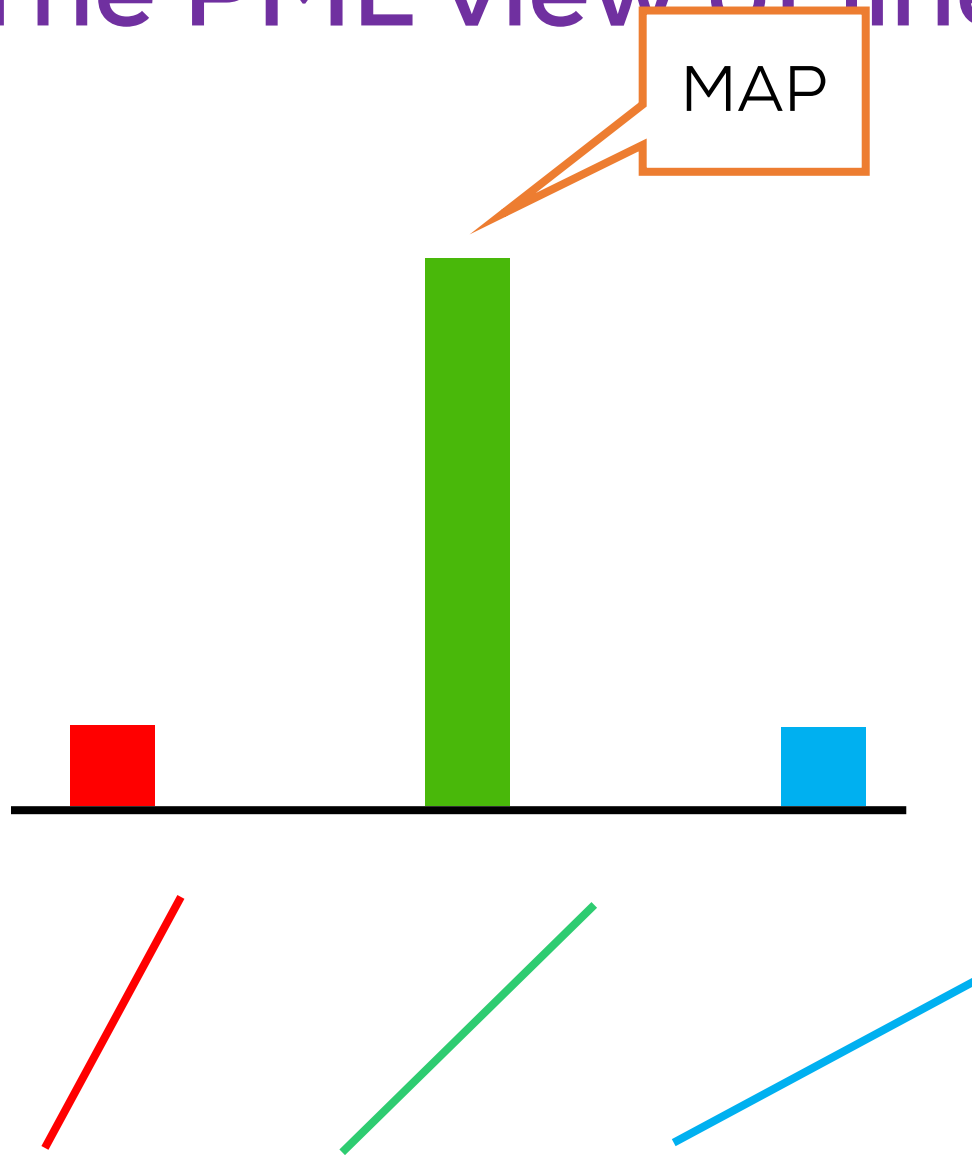
The PML view of linear regression



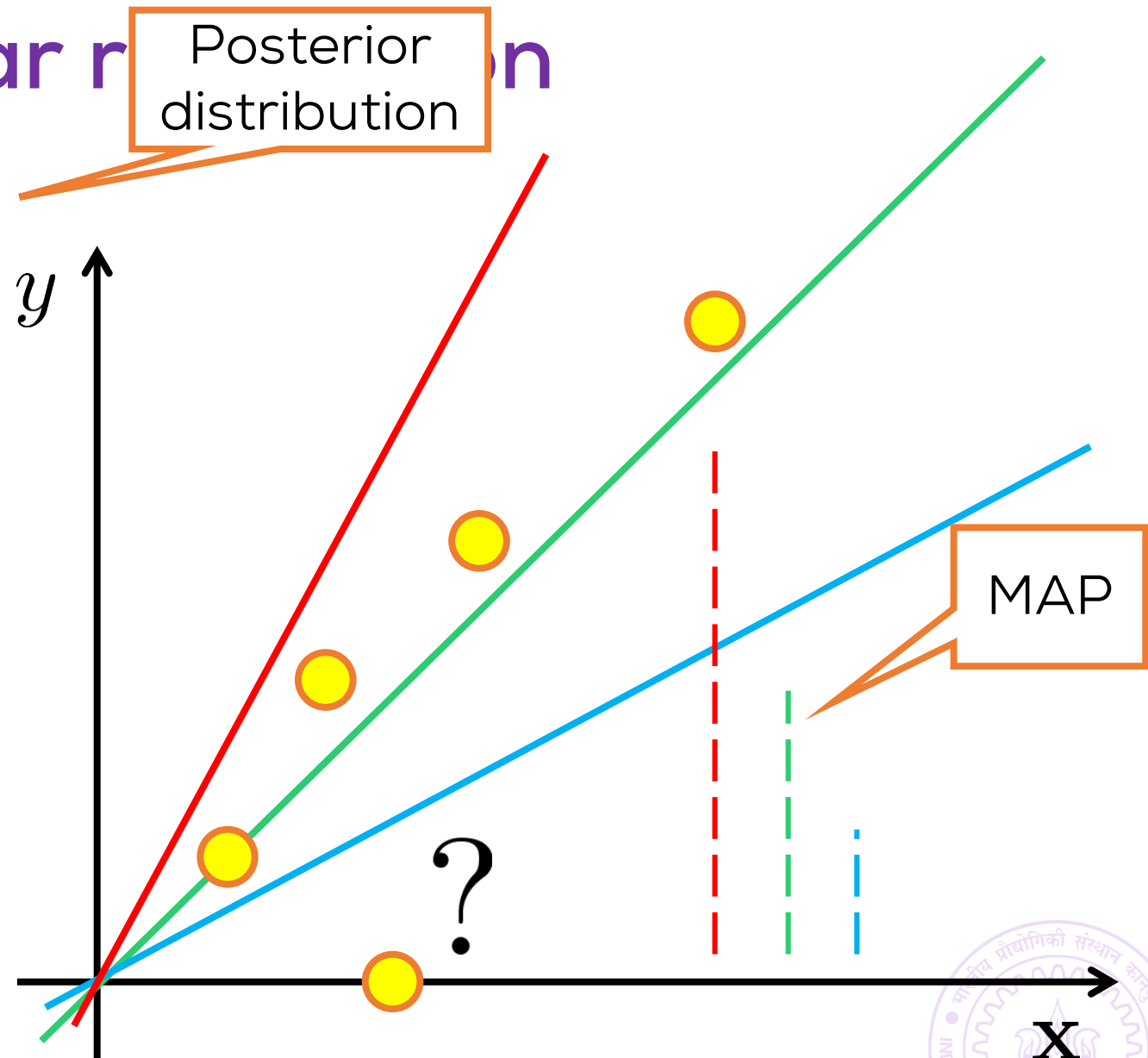
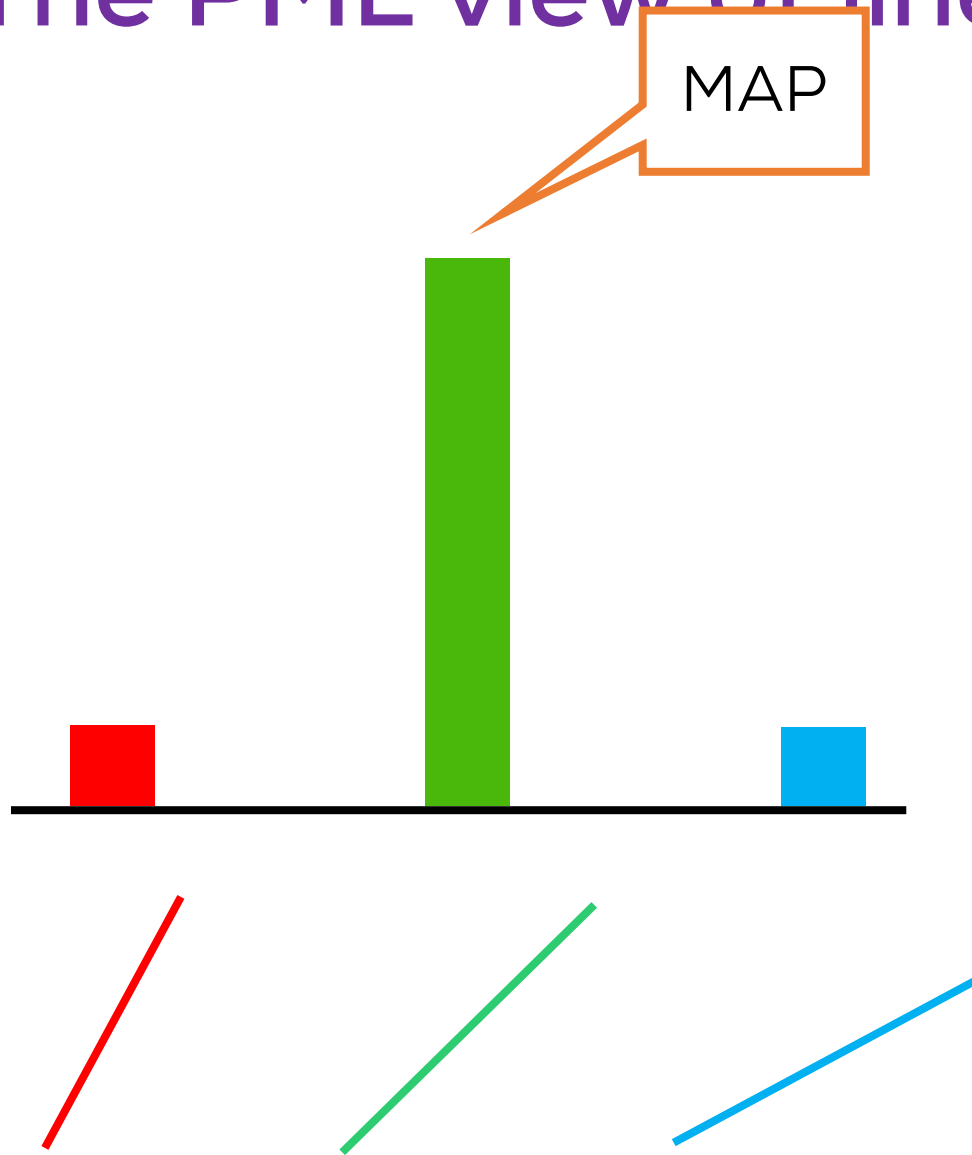
The PML view of linear regression



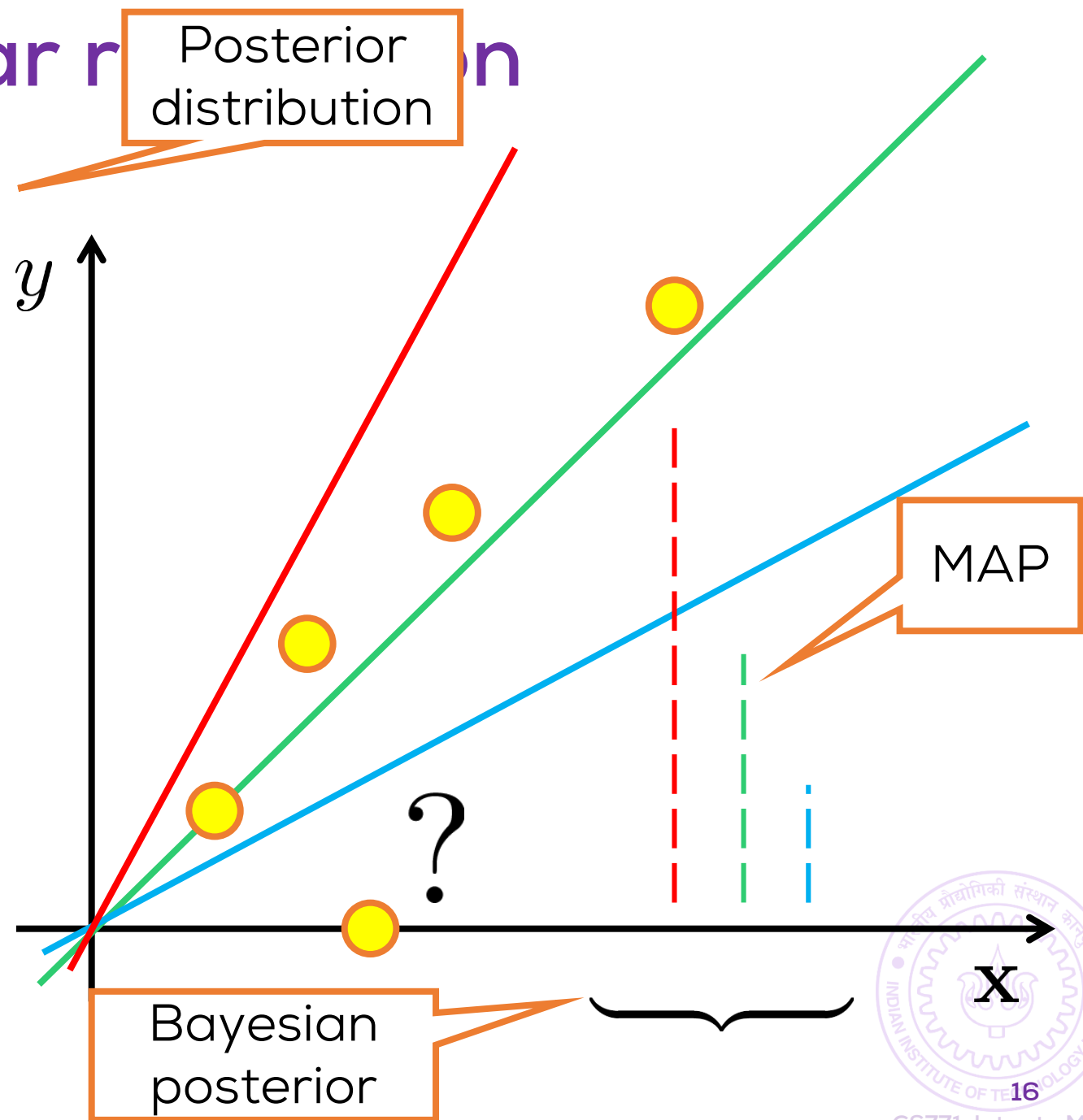
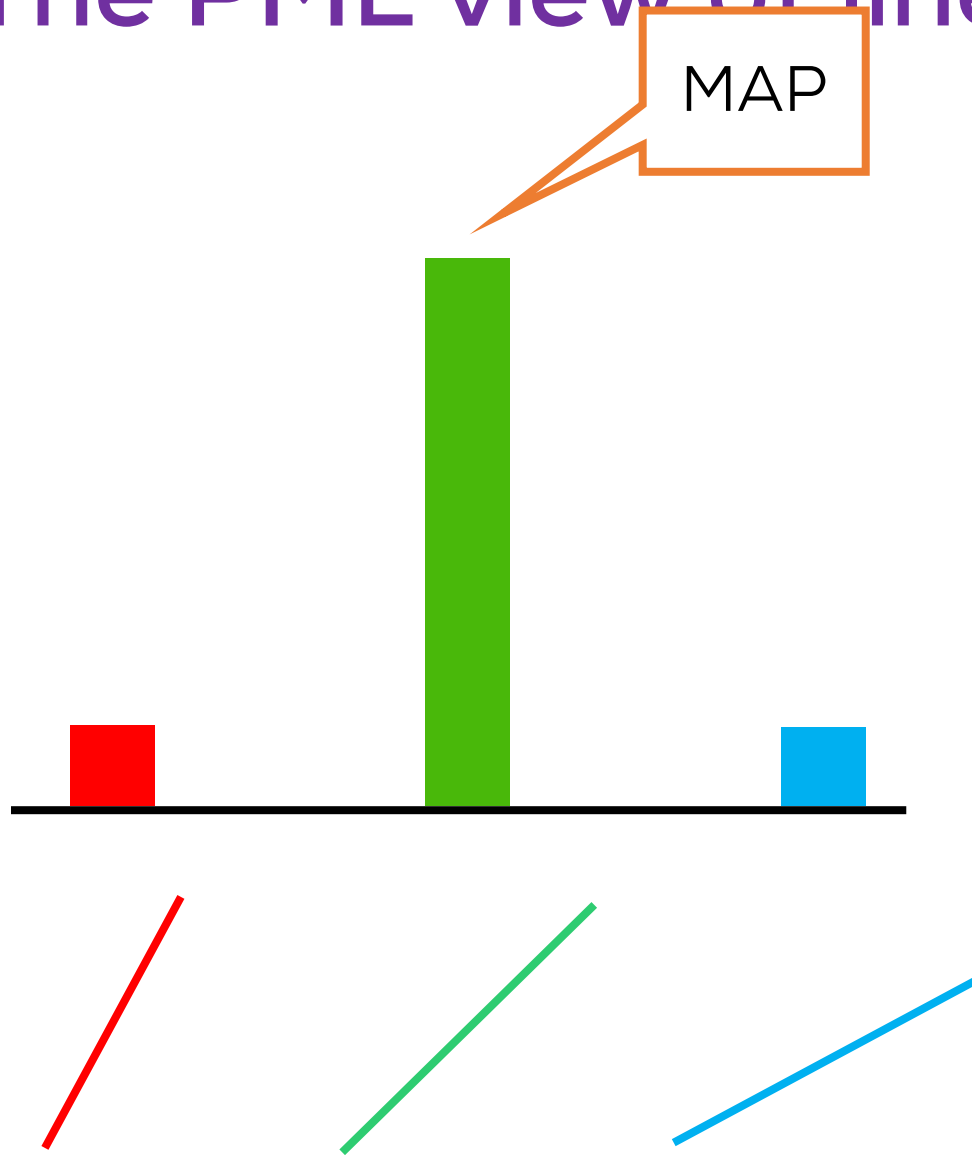
The PML view of linear regression



The PML view of linear regression



The PML view of linear regression



Linear Regression using MLE

$$\mathbb{P} [y \mid \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} \right)$$

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear
function

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear function

Likelihood

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear function

Likelihood

$$\log \mathbb{P}[\mathbf{y} | \mathbf{X}, \mathbf{w}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear function

Likelihood

$$\log \mathbb{P}[\mathbf{y} | \mathbf{X}, \mathbf{w}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

Log-likelihood

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear function

Likelihood

$$\log \mathbb{P}[\mathbf{y} | \mathbf{X}, \mathbf{w}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

Log-likelihood

$$\hat{\mathbf{w}}_{\text{MLE}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear function

Likelihood

$$\log \mathbb{P}[\mathbf{y} | \mathbf{X}, \mathbf{w}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

Log-likelihood

$$\hat{\mathbf{w}}_{\text{MLE}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 = (\mathbf{X}\mathbf{X}^\top)^\dagger \mathbf{X}\mathbf{y}$$

Linear Regression using MLE

$$\mathbb{P}[y | \mathbf{x}^i, \mathbf{w}] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear
function

Likelihood

$$\log \mathbb{P}[\mathbf{y} | \mathbf{X}, \mathbf{w}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

Log-likelihood

Least Squares!

$$\hat{\mathbf{w}}_{\text{MLE}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 = (\mathbf{X}\mathbf{X}^\top)^\dagger \mathbf{X}\mathbf{y}$$

Linear Regression using MAP

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2}\right)$$

Linear Regression using MAP

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp \left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2} \right)$$

$$\log \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 - \frac{1}{2\rho^2} \|\mathbf{w}\|_2^2$$

Linear Regression using MAP

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2}\right)$$

$$\log \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 - \frac{1}{2\rho^2} \|\mathbf{w}\|_2^2$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \frac{\sigma^2}{\rho^2} \|\mathbf{w}\|_2^2$$

Linear Regression using MAP

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2}\right)$$

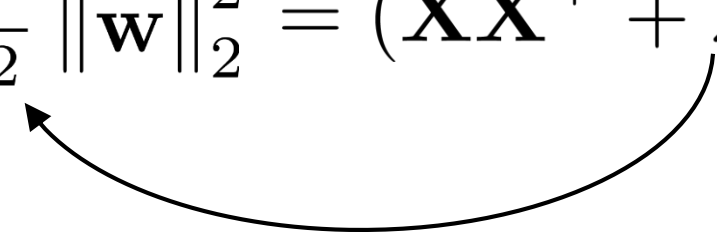
$$\log \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 - \frac{1}{2\rho^2} \|\mathbf{w}\|_2^2$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \frac{\sigma^2}{\rho^2} \|\mathbf{w}\|_2^2 = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1} \mathbf{X}\mathbf{y}$$

Linear Regression using MAP

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2}\right)$$

$$\log \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 - \frac{1}{2\rho^2} \|\mathbf{w}\|_2^2$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \frac{\sigma^2}{\rho^2} \|\mathbf{w}\|_2^2 = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1} \mathbf{X}\mathbf{y}$$


Linear Regression using MAP

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2}\right)$$

$$\log \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 - \frac{1}{2\rho^2} \|\mathbf{w}\|_2^2$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg \min \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \frac{\sigma^2}{\rho^2} \|\mathbf{w}\|_2^2 = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1} \mathbf{X}\mathbf{y}$$

Ridge Regression

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior
is Gaussian like
the prior!

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior
is Gaussian like
the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$

Wait! MAP?

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$

Wait! MAP?

By definition, MAP is the mode of the posterior

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$
$$\Lambda = \left(\frac{1}{\sigma^2} \mathbf{X}\mathbf{X}^\top + \frac{1}{\rho^2} \cdot I \right)^{-1}$$

Wait! MAP?

By definition, MAP is the mode of the posterior

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$
$$\Lambda = \left(\frac{1}{\sigma^2} \mathbf{X}\mathbf{X}^\top + \frac{1}{\rho^2} \cdot I \right)^{-1}$$

Wait! MAP?

By definition, MAP is the mode of the posterior

$$\mathbb{P}[y \mid \mathbf{x}, \mathbf{X}, \mathbf{y}] = \int_{\mathbf{w}} \mathbb{P}[y \mid \mathbf{x}, \mathbf{w}] \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] d\mathbf{w}$$

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$
$$\Lambda = \left(\frac{1}{\sigma^2} \mathbf{X}\mathbf{X}^\top + \frac{1}{\rho^2} \cdot I \right)^{-1}$$

Wait! MAP?

By definition, MAP is the mode of the posterior

$$\mathbb{P}[y \mid \mathbf{x}, \mathbf{X}, \mathbf{y}] = \int_{\mathbf{w}} \mathbb{P}[y \mid \mathbf{x}, \mathbf{w}] \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] d\mathbf{w}$$
$$= \mathcal{N}(\langle \boldsymbol{\nu}, \mathbf{x} \rangle, \sigma^2 + \mathbf{x}^\top \Lambda \mathbf{x})$$

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$
$$\Lambda = \left(\frac{1}{\sigma^2} \mathbf{X}\mathbf{X}^\top + \frac{1}{\rho^2} \cdot I \right)^{-1}$$

Wait! MAP?

By definition, MAP is the mode of the posterior

$$\mathbb{P}[y \mid \mathbf{x}, \mathbf{X}, \mathbf{y}] = \int_{\mathbf{w}} \mathbb{P}[y \mid \mathbf{x}, \mathbf{w}] \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] d\mathbf{w}$$
$$= \mathcal{N}(\langle \boldsymbol{\nu}, \mathbf{x} \rangle, \sigma^2 + \mathbf{x}^\top \Lambda \mathbf{x})$$

Predictive Posterior

Bayesian Linear Regression?

$$\mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2} \cdot I \right)^{-1} \mathbf{X}\mathbf{y}$$
$$\Lambda = \left(\frac{1}{\sigma^2} \mathbf{X}\mathbf{X}^\top + \frac{1}{\rho^2} \cdot I \right)^{-1}$$

Wait! MAP?

By definition, MAP is the mode of the posterior

$$\mathbb{P}[y \mid \mathbf{x}, \mathbf{X}, \mathbf{y}] = \int_{\mathbf{w}} \mathbb{P}[y \mid \mathbf{x}, \mathbf{w}] \mathbb{P}[\mathbf{w} \mid \mathbf{X}, \mathbf{y}] d\mathbf{w}$$
$$= \mathcal{N}(\langle \boldsymbol{\nu}, \mathbf{x} \rangle, \sigma^2 + \mathbf{x}^\top \Lambda \mathbf{x})$$

Predictive Posterior

Extra Information

Function Approximation

Sept 3, 2017



FA Queries

- *Why does regularization help avoid overfitting?*

We have already seen an example in lecture where unregularized models could try to fit noise and give poor result.

- *What are "sparse" regularizers? Why does anything being sparse help?*

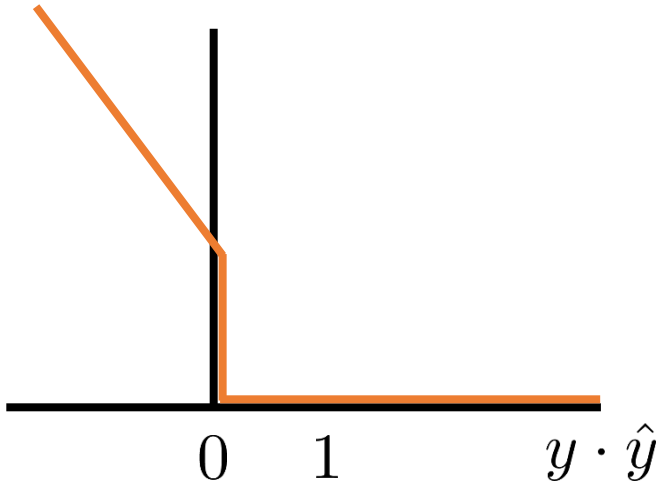
Regularizers such as L_1 norm $\|\mathbf{w}\|_1 = \sum_{j=1}^d |\mathbf{w}_j|$ and entropy $\sum_{j=1}^d \mathbf{w}_j \log \mathbf{w}_j$ are known to often lead to solutions that are sparse i.e. $\|\mathbf{w}\|_0 = |\{j: \mathbf{w}_j \neq 0\}|$ is small. See Piazza post for more. Sparse models take less storage as well as they are faster at prediction (computing $\langle \mathbf{w}, \mathbf{x} \rangle$ is faster if \mathbf{w} is sparse).

- *How do you decide the constraint in constrained optimization? That cant be inductive bias, right, as that would mean that even before seeing the data you have a very good estimate of what the curve should look like.*

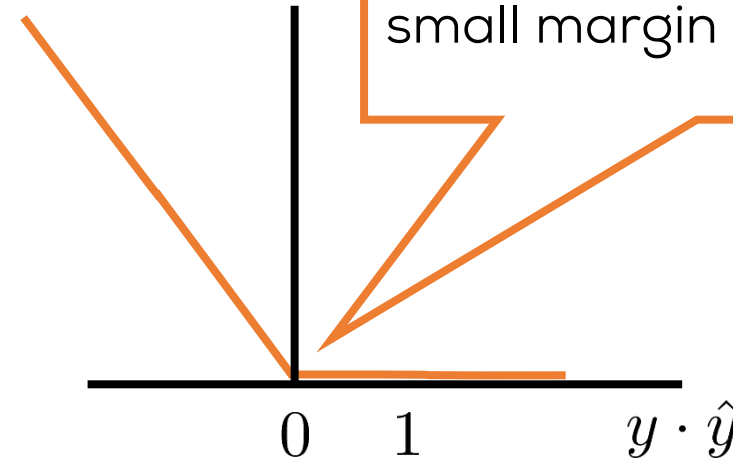
Not a very good estimate but a rough idea of what it should be. For example in the constraint $\|\mathbf{w}\|_2 \leq r$, the form is decided before but not r

FA Queries

- *Why is the truncated hinge loss graph the way it is?*



Truncated Hinge loss

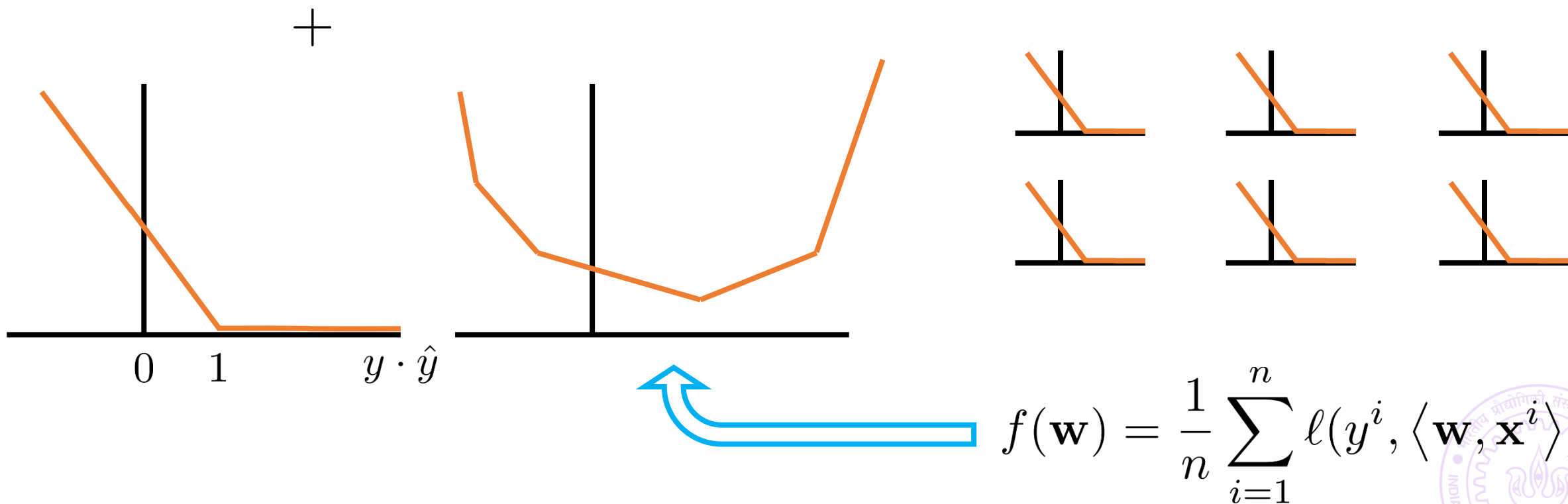


New Truncated Hinge loss

Assigns very small loss to misclassifications with a small margin – very forgiving.

FA Queries

- For a convex function, gradient vanishes at local minima. Even if there is a kink at that point and the function is not differentiable, how does it matter? Since the function is convex, we can conclude that we have found the minima. Why bother with sub gradient and sub differentials etc?



Multi-classification Loss Functions

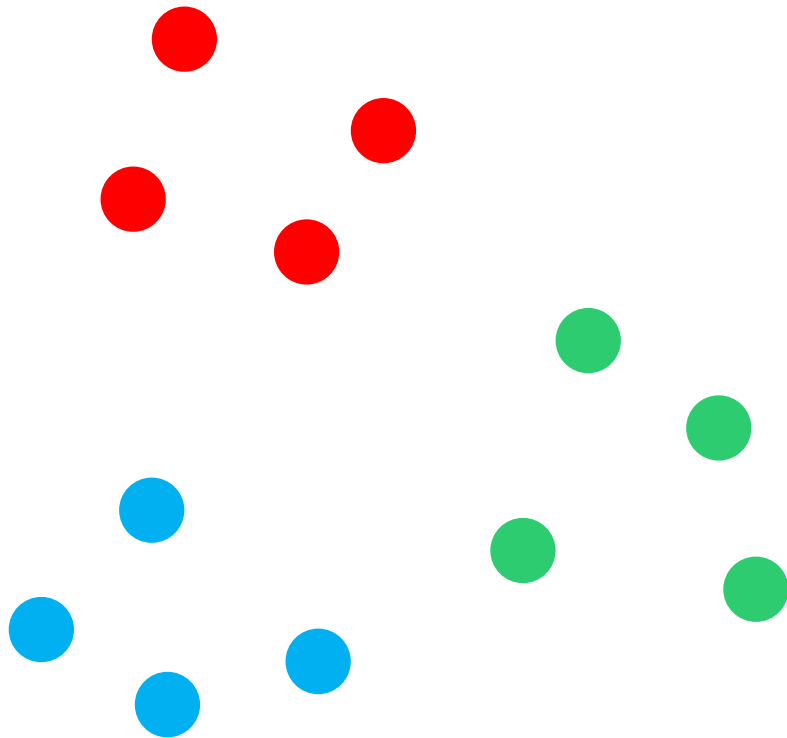
One-vs-All (OVA)

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

Multi-classification Loss Functions

One-vs-All (OVA)

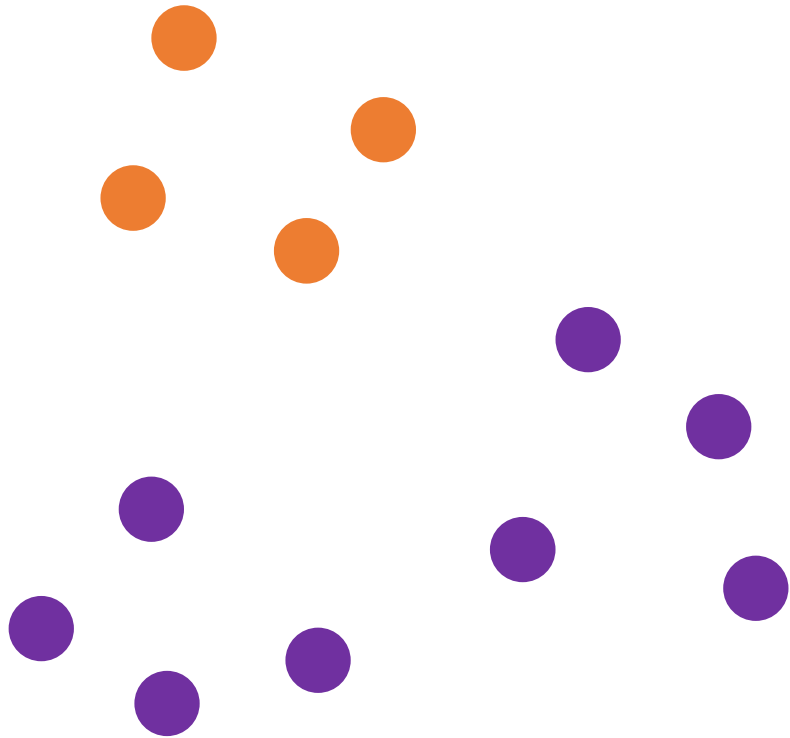
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

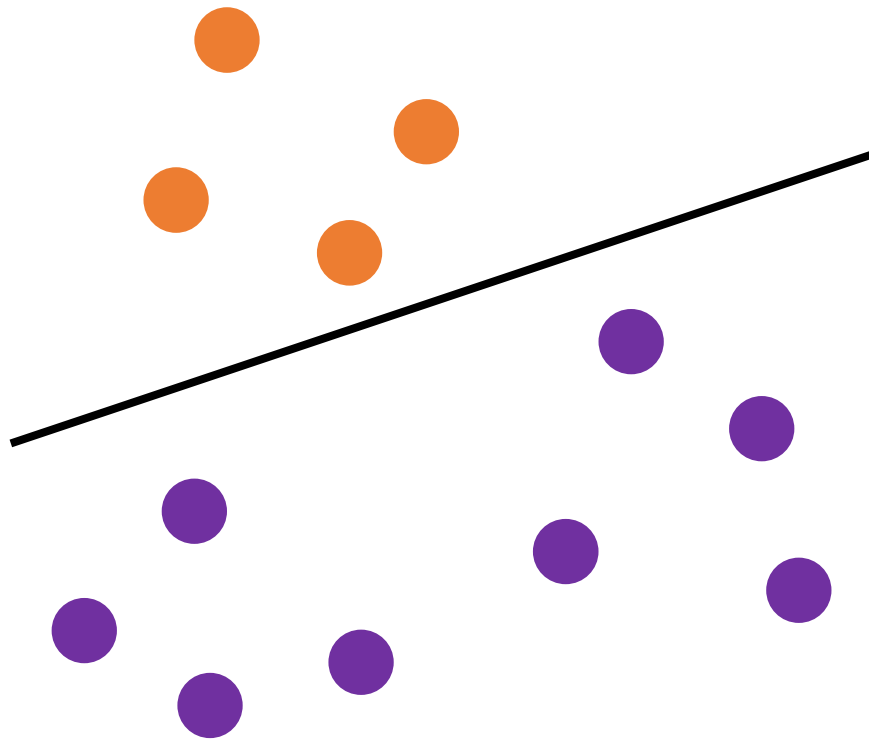
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

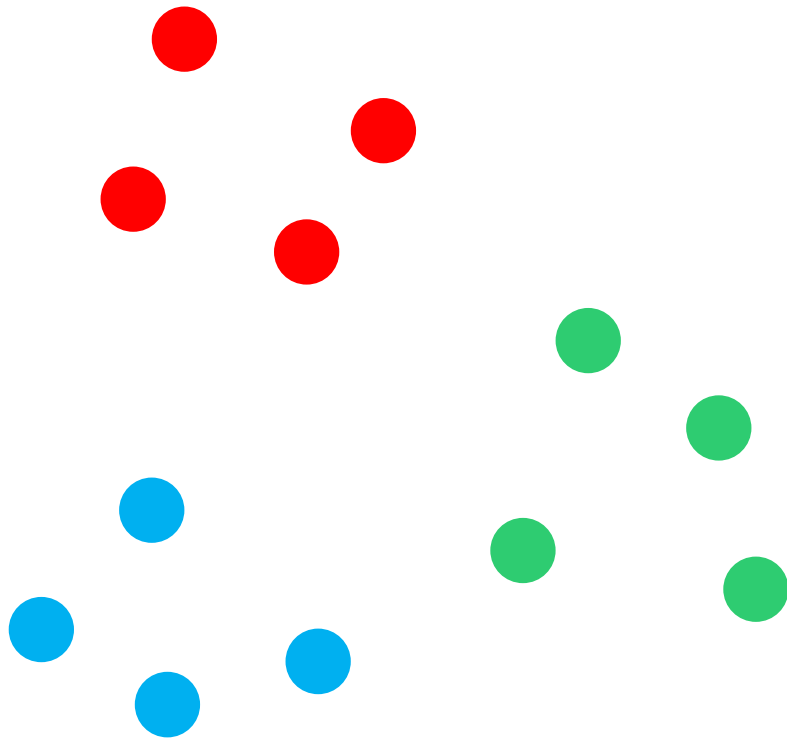
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

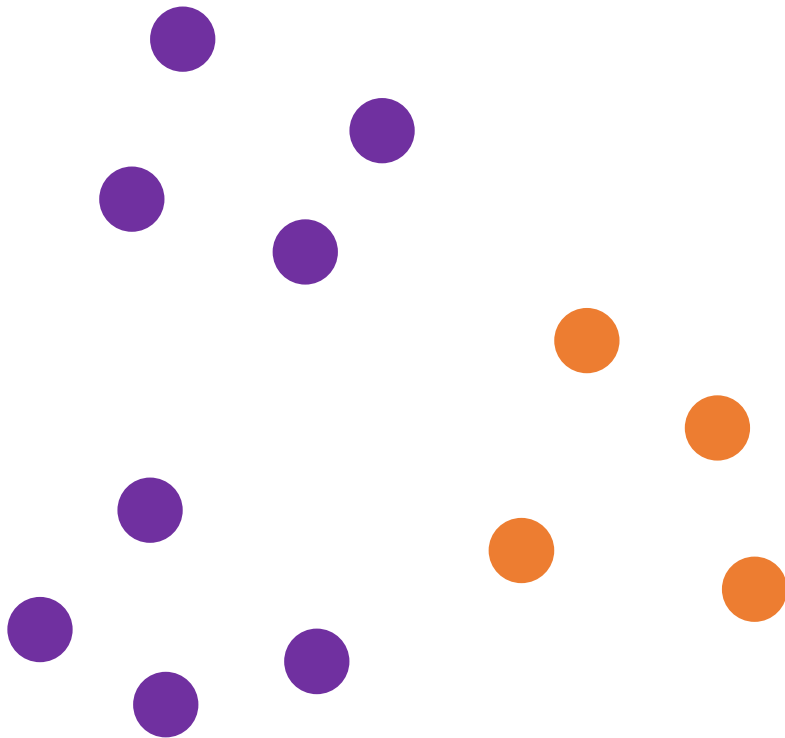
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

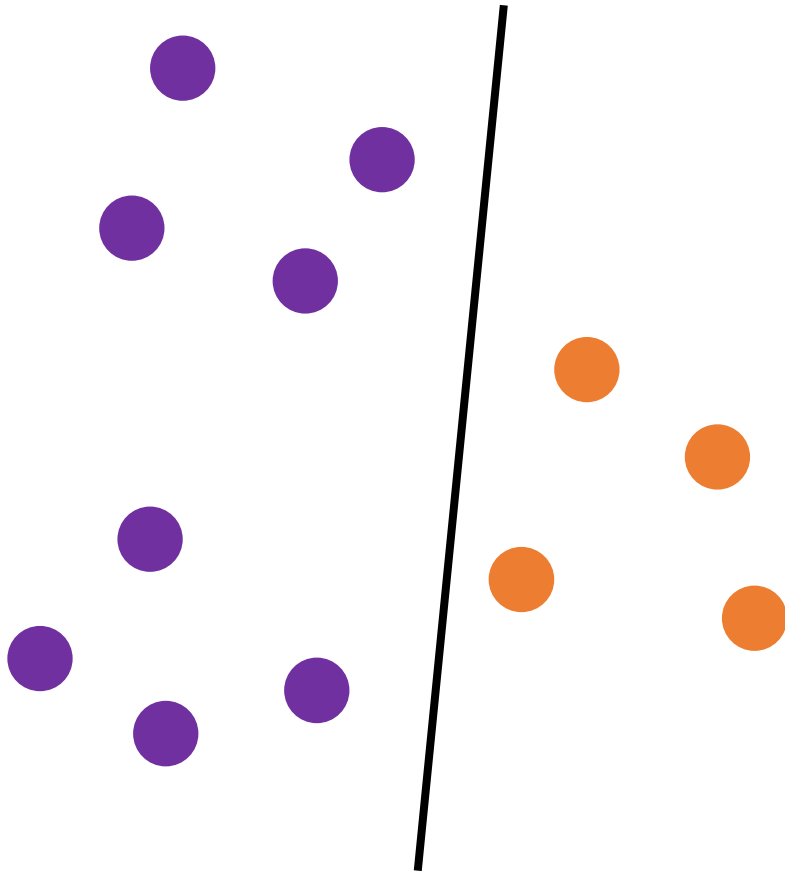
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

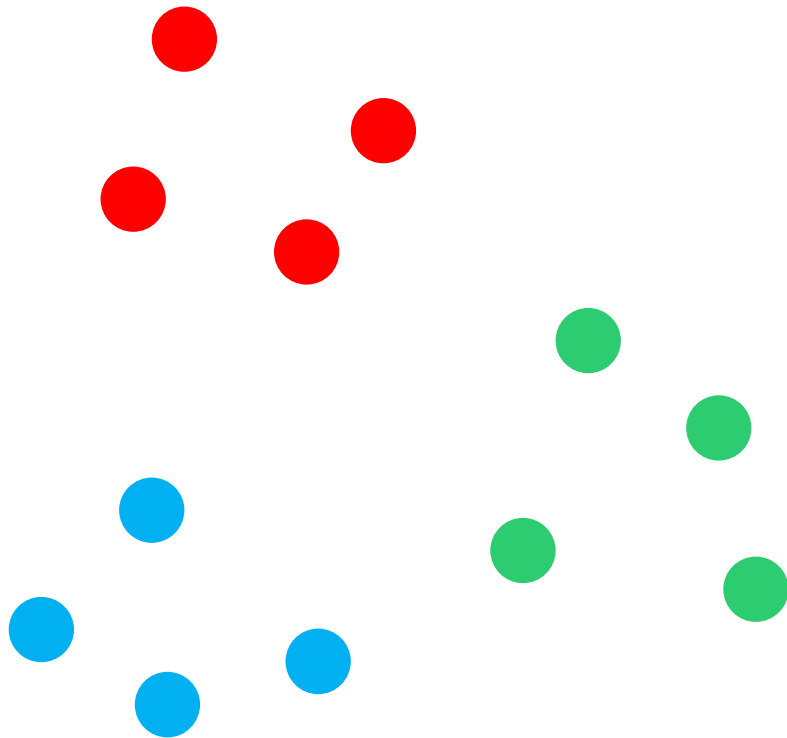
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

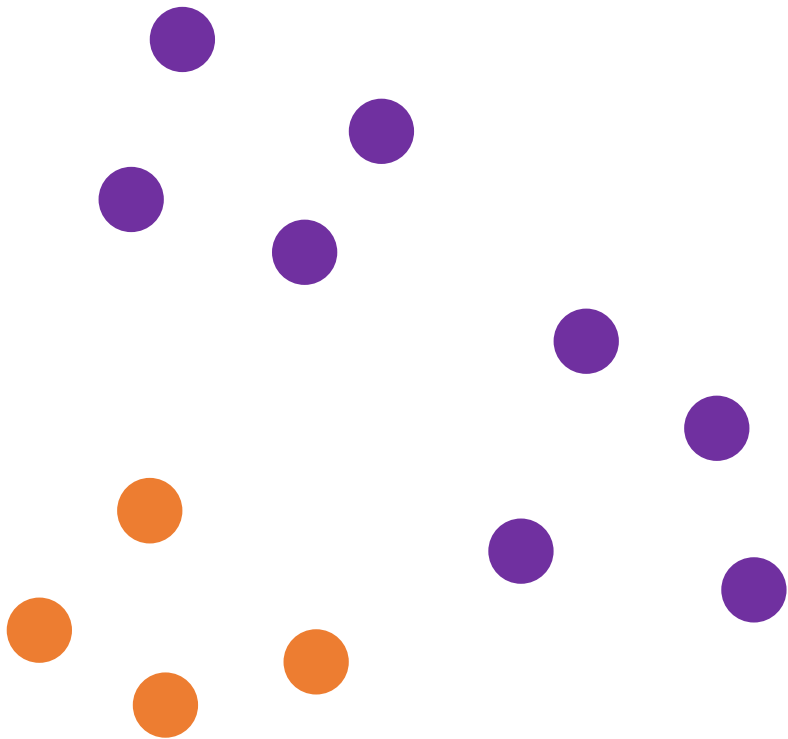
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

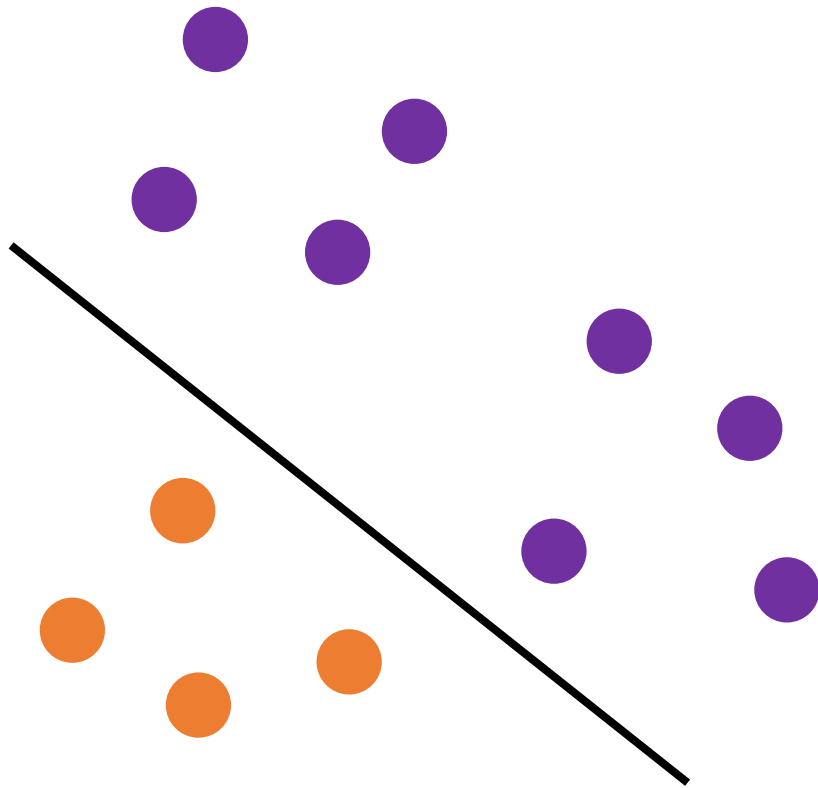
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

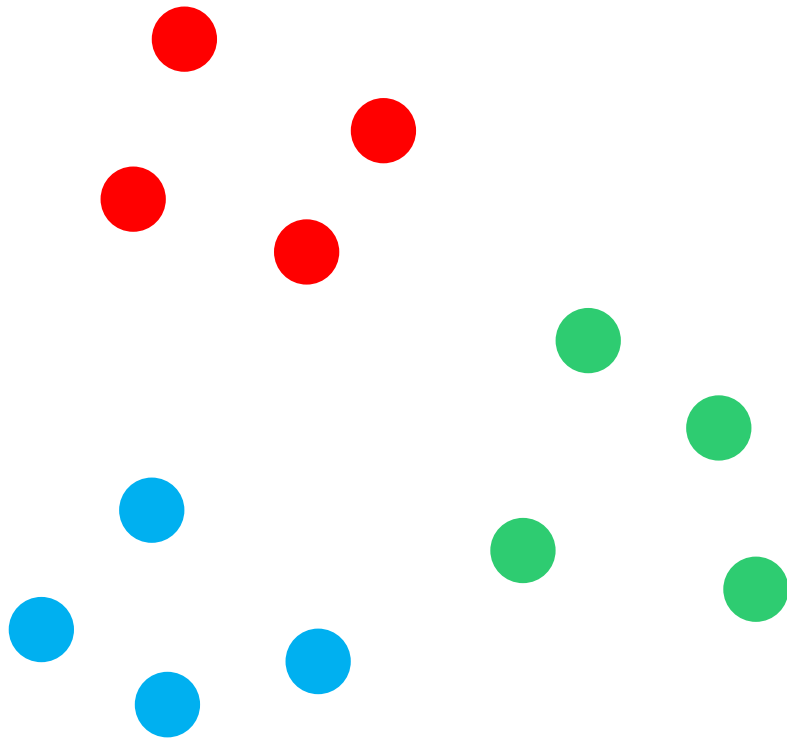
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

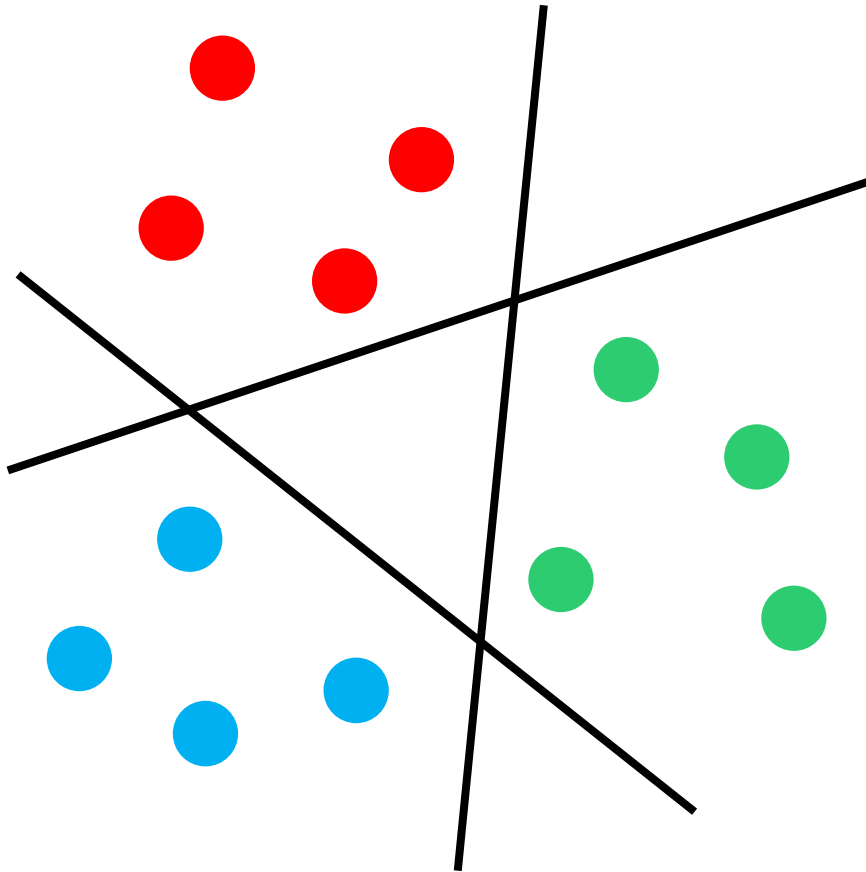
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

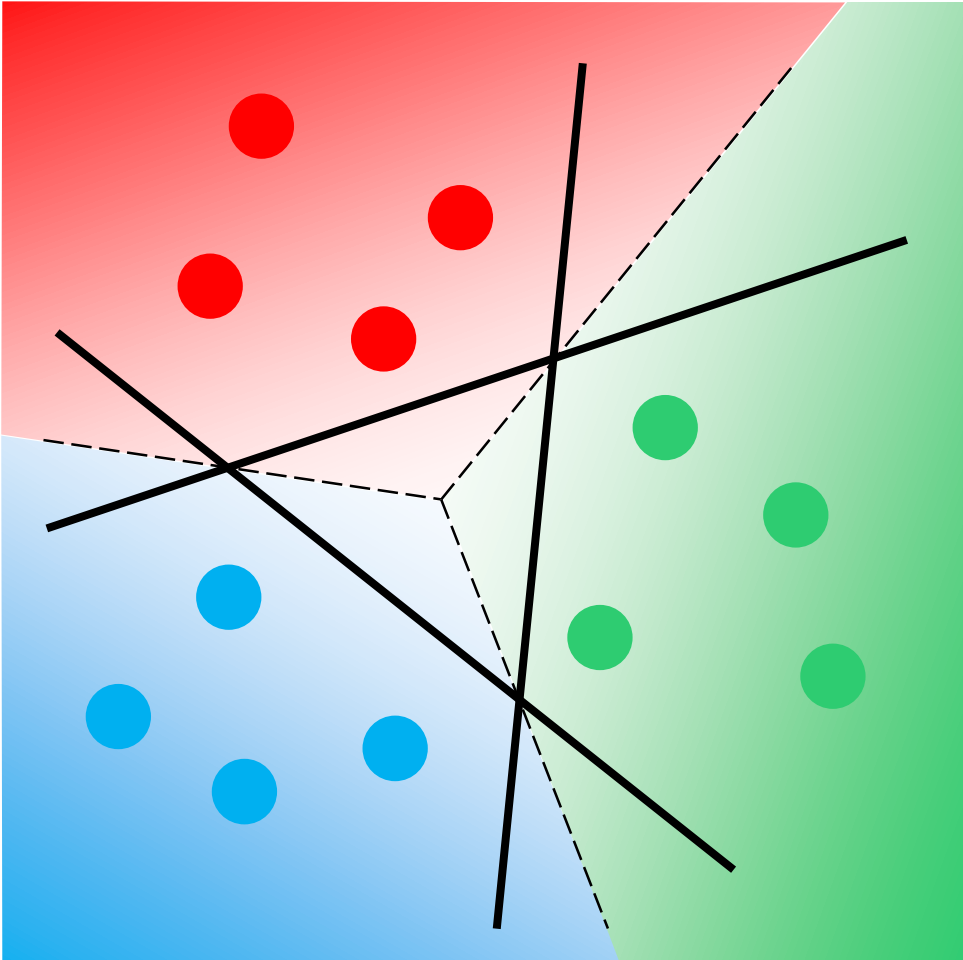
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

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

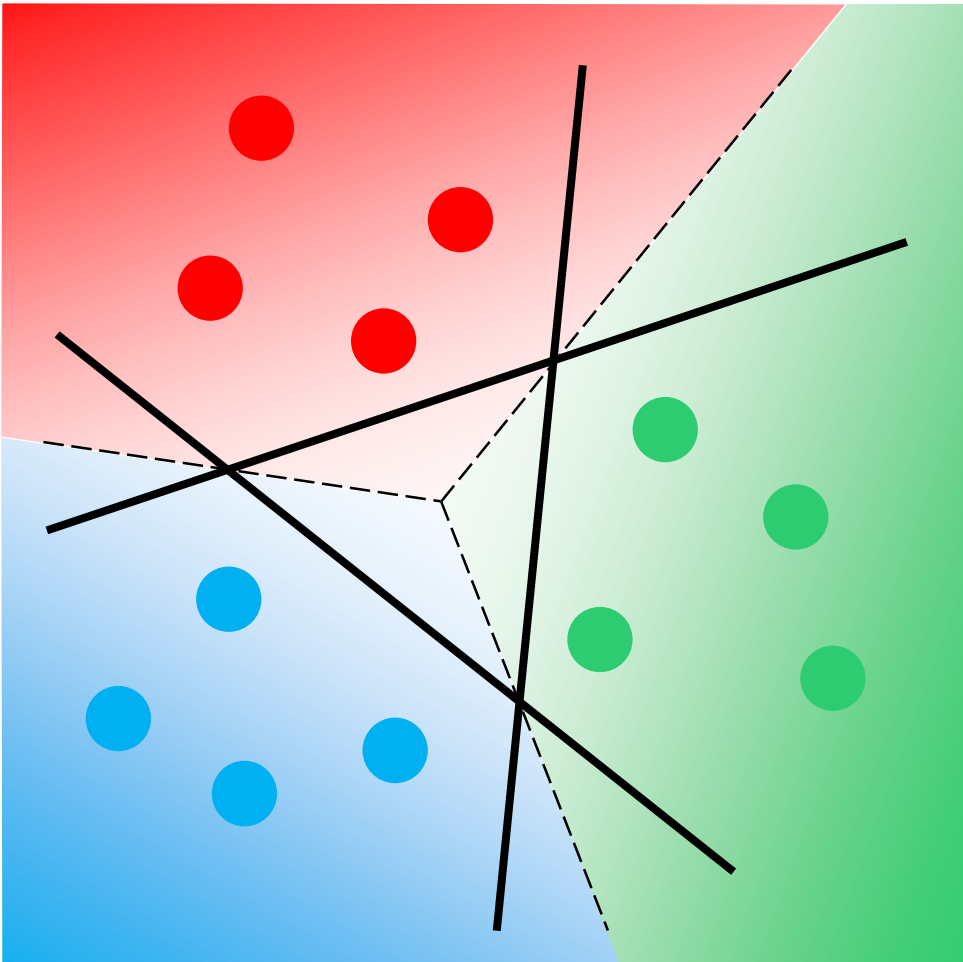


Multi-classification Loss Functions

One-vs-All (OVA)

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

$$\hat{\mathbf{w}}^j = \arg \min_{\mathbf{w}} \sum_{i=1}^n \ell(y^{i,(j)}, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$
$$y^{i,(j)} = \begin{cases} 1 & ; y^i = j \\ -1 & ; y^i \neq j \end{cases}$$



Multi-classification Loss Functions

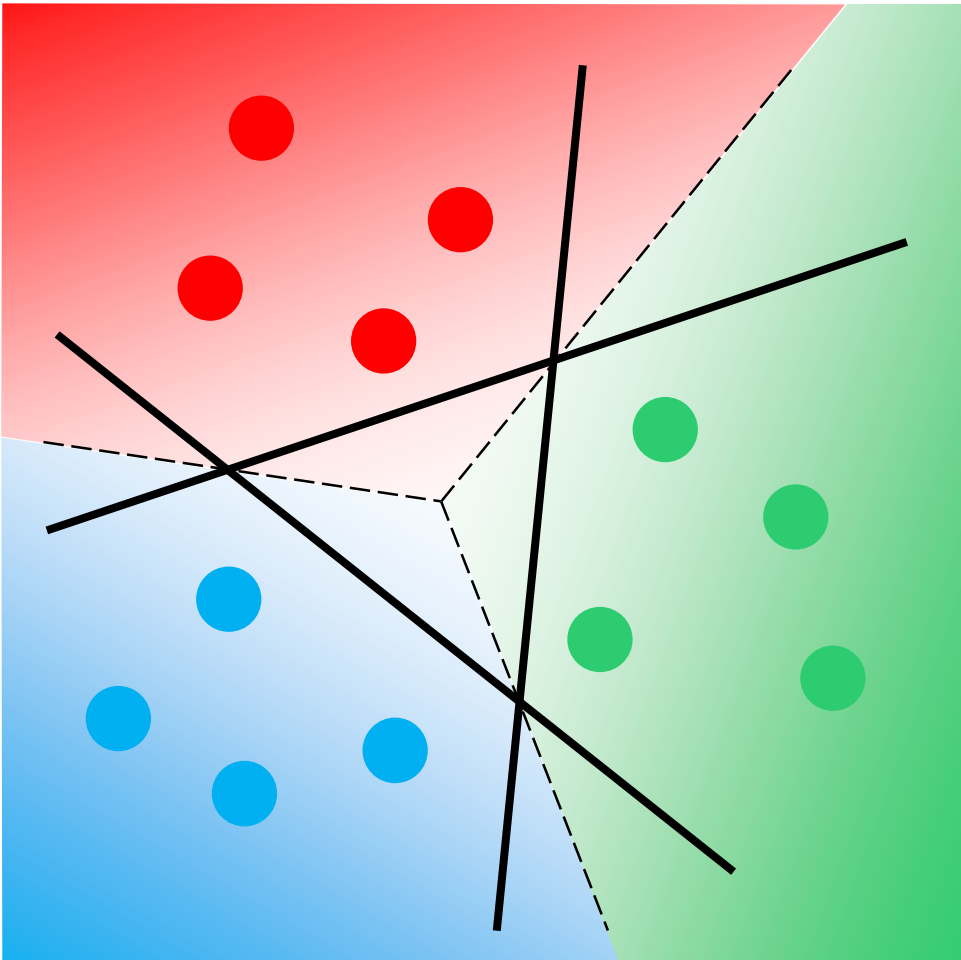
One-vs-All (OVA)

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

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

hinge, logistic etc

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



Multi-classification Loss Functions

One-vs-All (OVA)

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

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

hinge, logistic etc

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

In OVA, after you get all the individual decision boundaries, how do you combine them into one solution?

Multi-classification Loss Functions

One-vs-All (OVA)

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

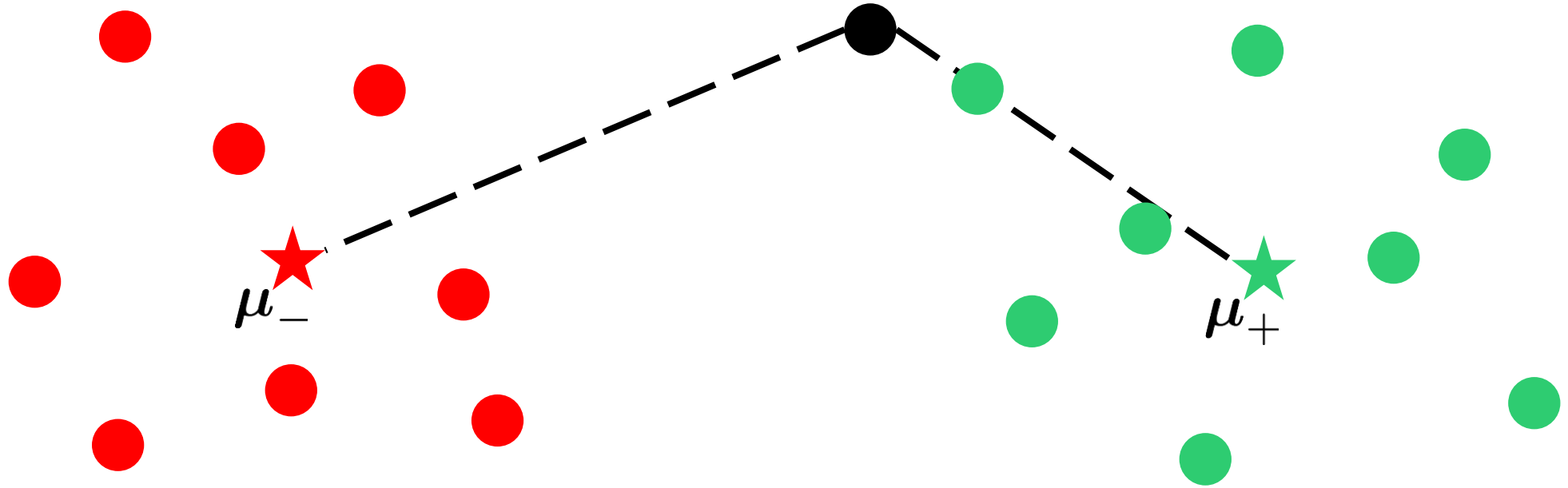
$$\hat{\mathbf{w}}^j = \arg \min_{\mathbf{w}} \sum_{i=1}^n \ell(y^{i,(j)}, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

hinge, logistic etc

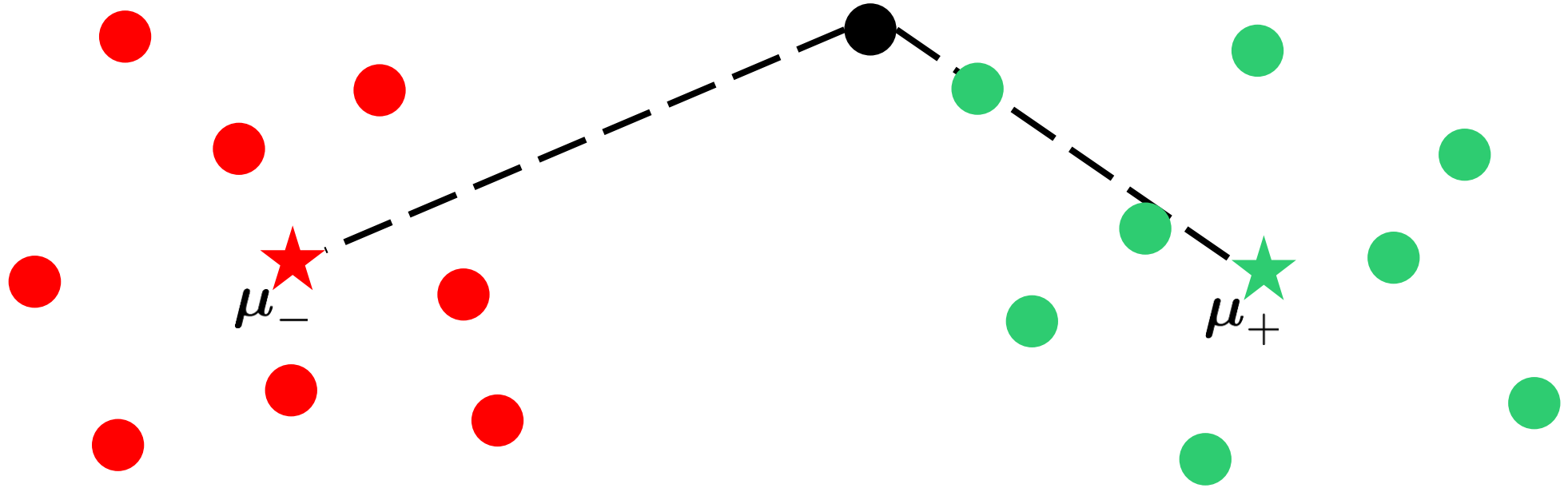
$$y^{i,(j)} = \begin{cases} 1 & ; y^i = j \\ -1 & ; y^i \neq j \end{cases}$$

In OVA, after you get all the individual decision boundaries, how do you combine them into one solution?

Learning with Prototypes

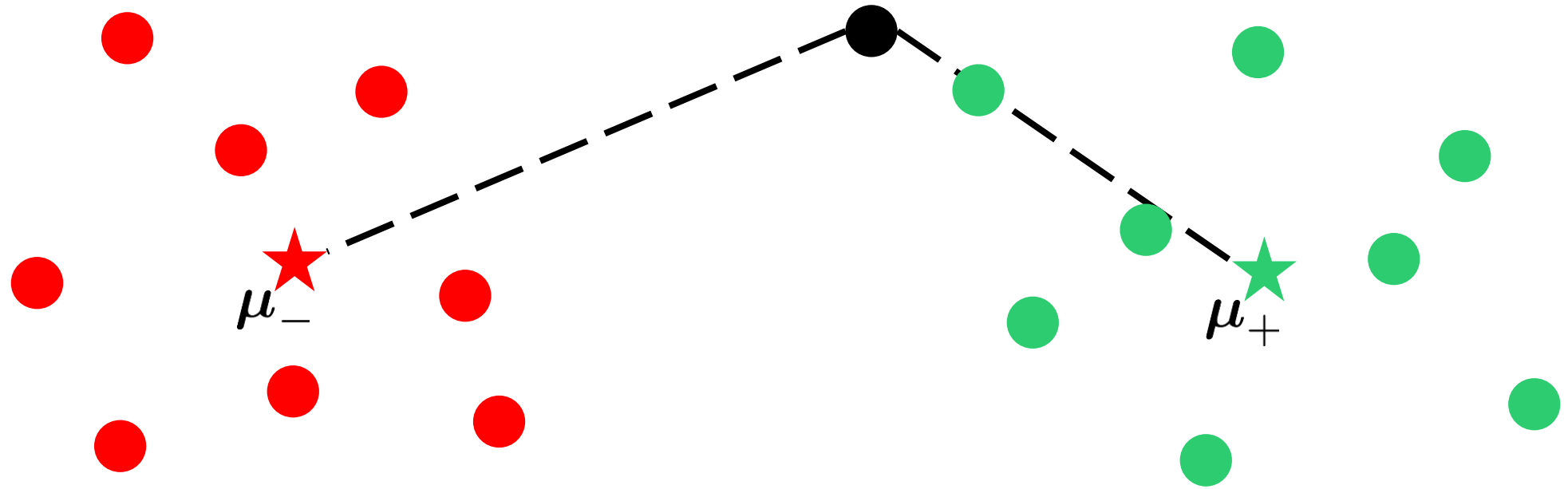


Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Learning with Prototypes

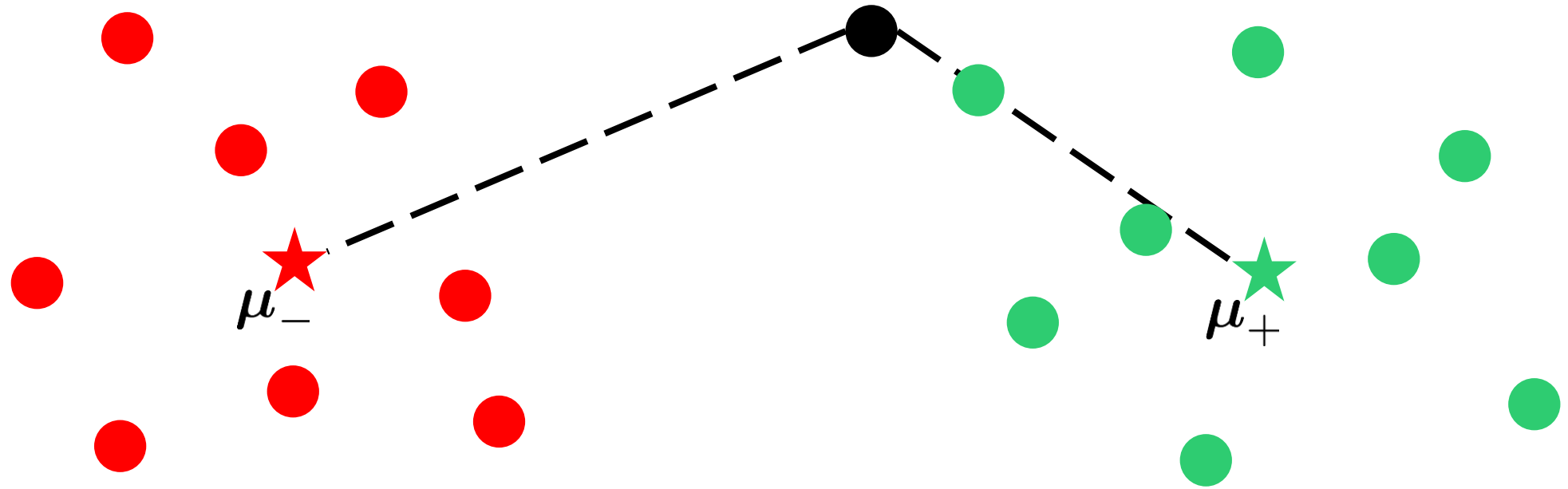


$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-)$$



Learning with Prototypes

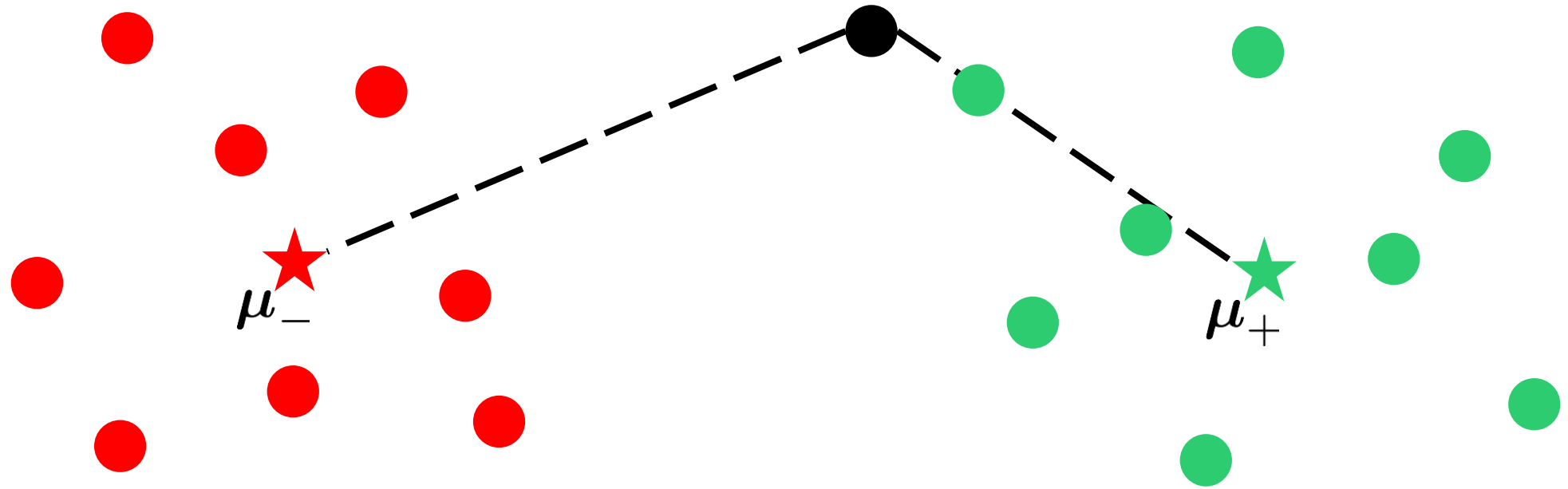


$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

Learning with Prototypes



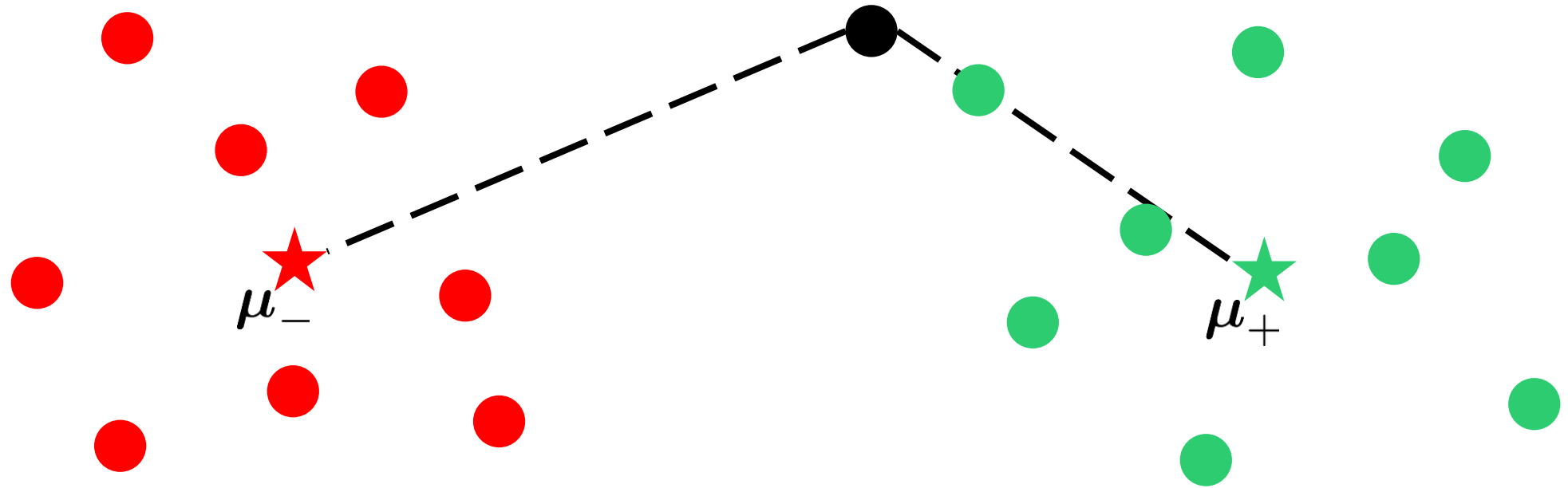
$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Decision Boundary

$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

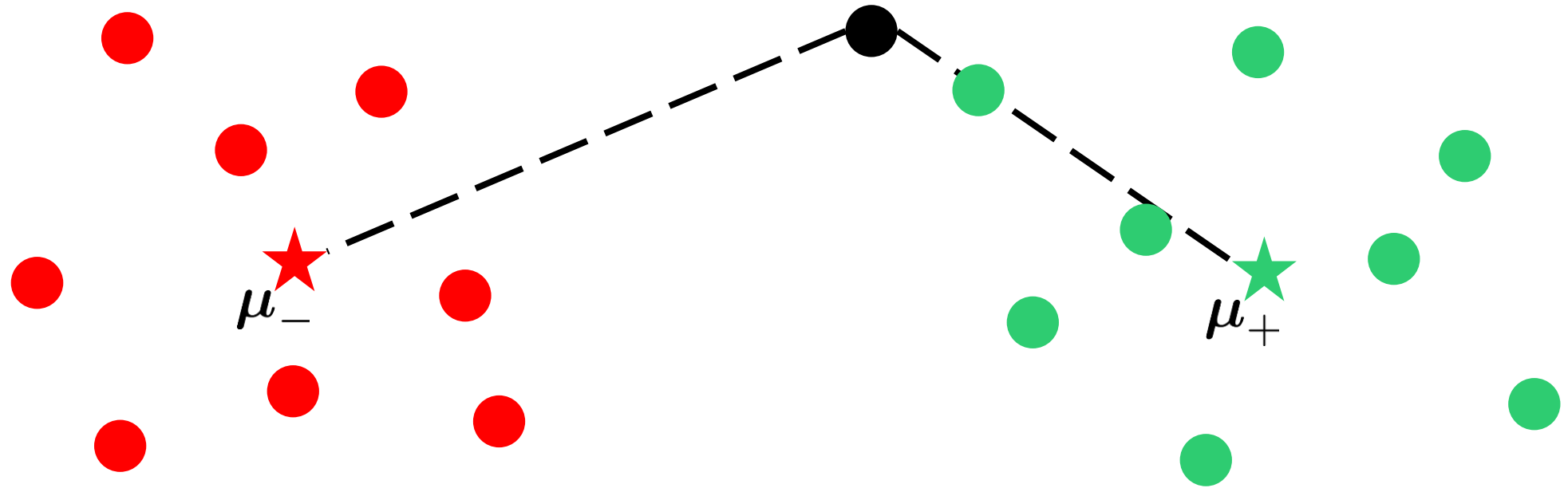
Decision Boundary

$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) = d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{black dot on boundary}$$

Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Decision Boundary

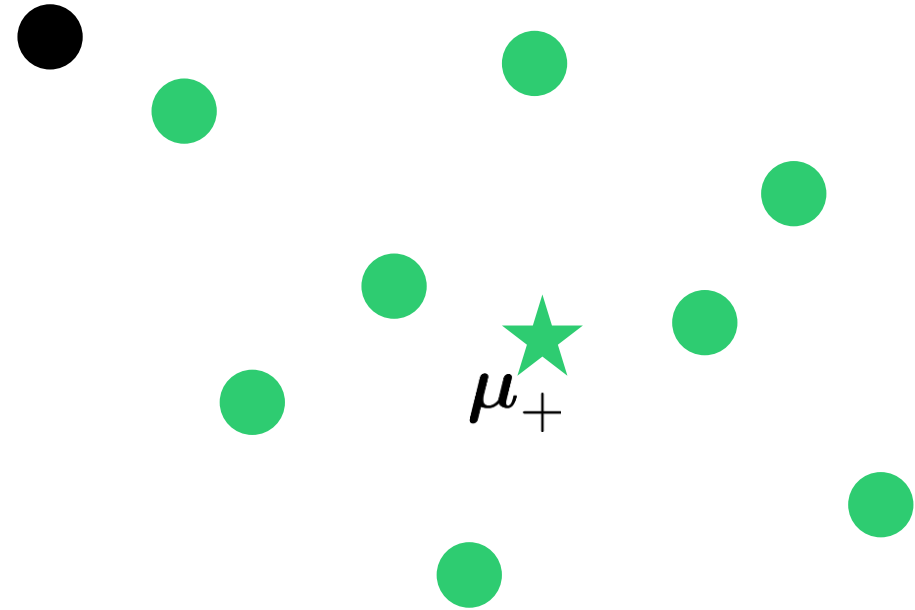
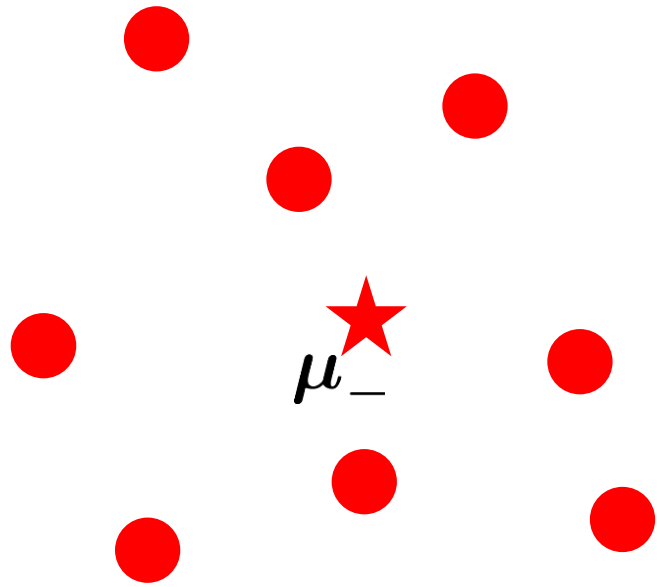
$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) = d(\mathbf{x}, \boldsymbol{\mu}_-)$$

$$\equiv \langle \mathbf{w}, \mathbf{x} \rangle + b = 0$$

Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Linear Decision Boundary

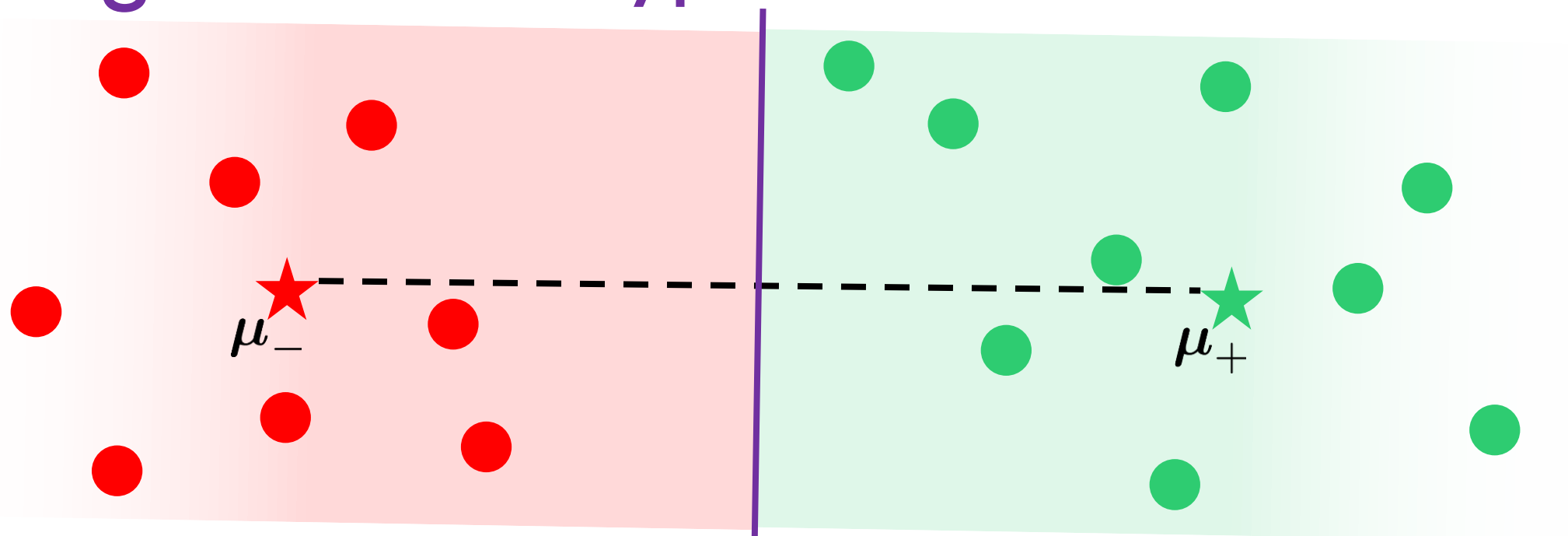
$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) = d(\mathbf{x}, \boldsymbol{\mu}_-)$$

$$\equiv \langle \mathbf{w}, \mathbf{x} \rangle + b = 0$$

Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Linear Decision Boundary

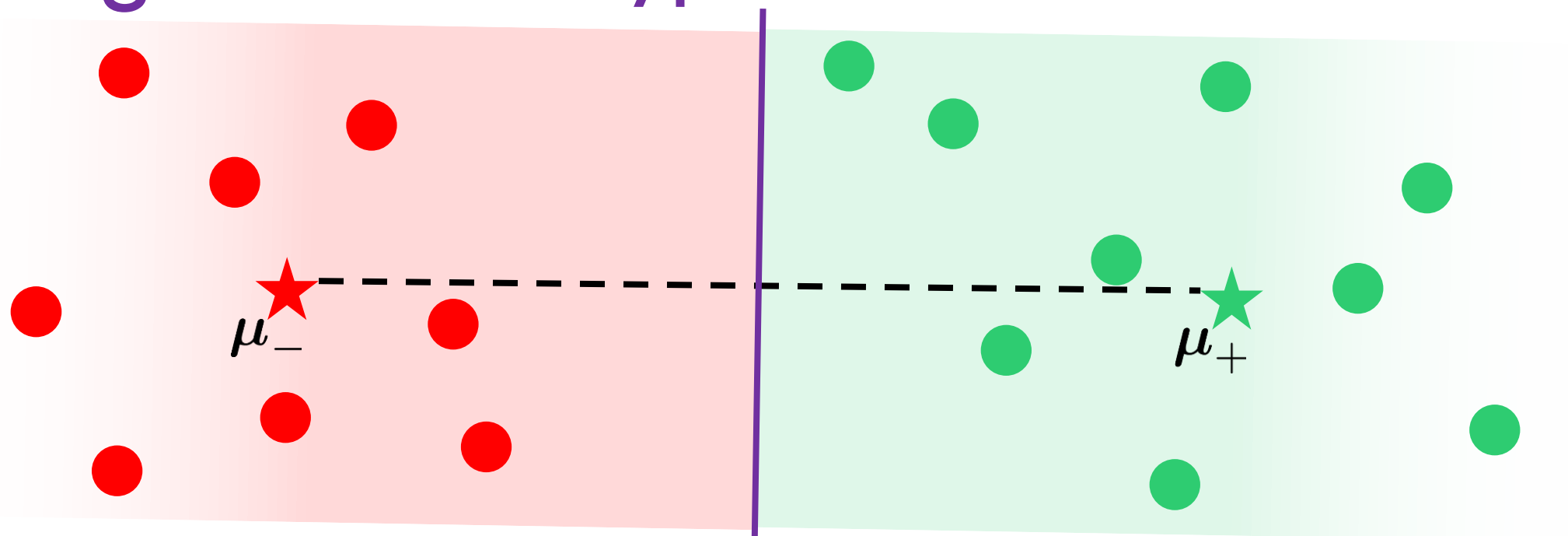
$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) = d(\mathbf{x}, \boldsymbol{\mu}_-)$$

$$\equiv \langle \mathbf{w}, \mathbf{x} \rangle + b = 0$$

Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Linear Decision Boundary

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

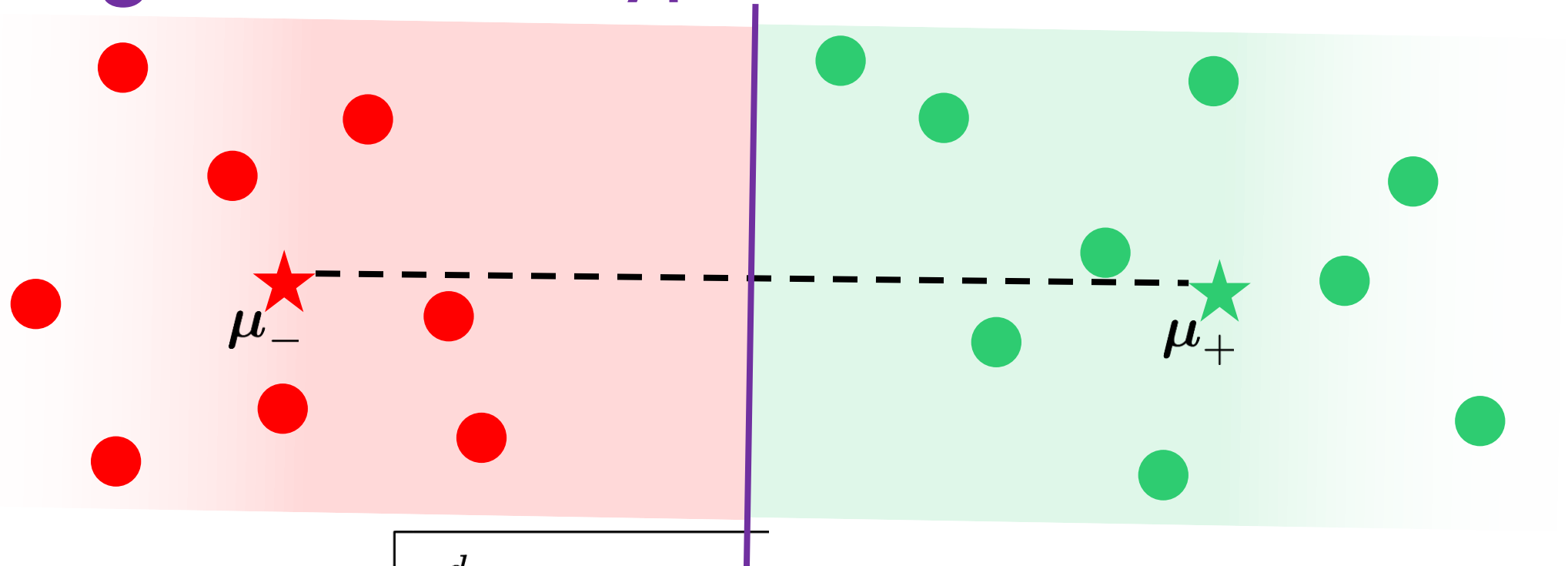
$$d(\mathbf{x}, \boldsymbol{\mu}_+) > d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{red circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) < d(\mathbf{x}, \boldsymbol{\mu}_-) \quad \text{green circle}$$

$$d(\mathbf{x}, \boldsymbol{\mu}_+) = d(\mathbf{x}, \boldsymbol{\mu}_-)$$

$$\equiv \langle \mathbf{w}, \mathbf{x} \rangle + b = 0$$

Learning with Prototypes

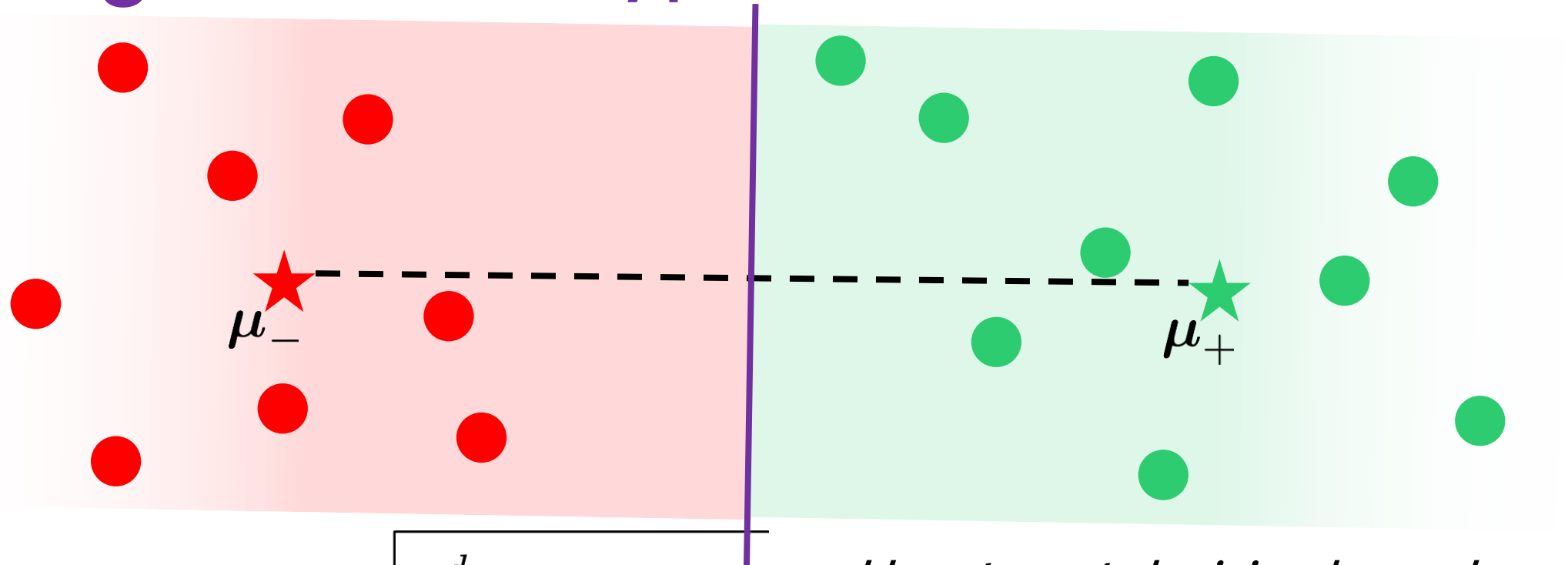


$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Linear Decision Boundary

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

Learning with Prototypes



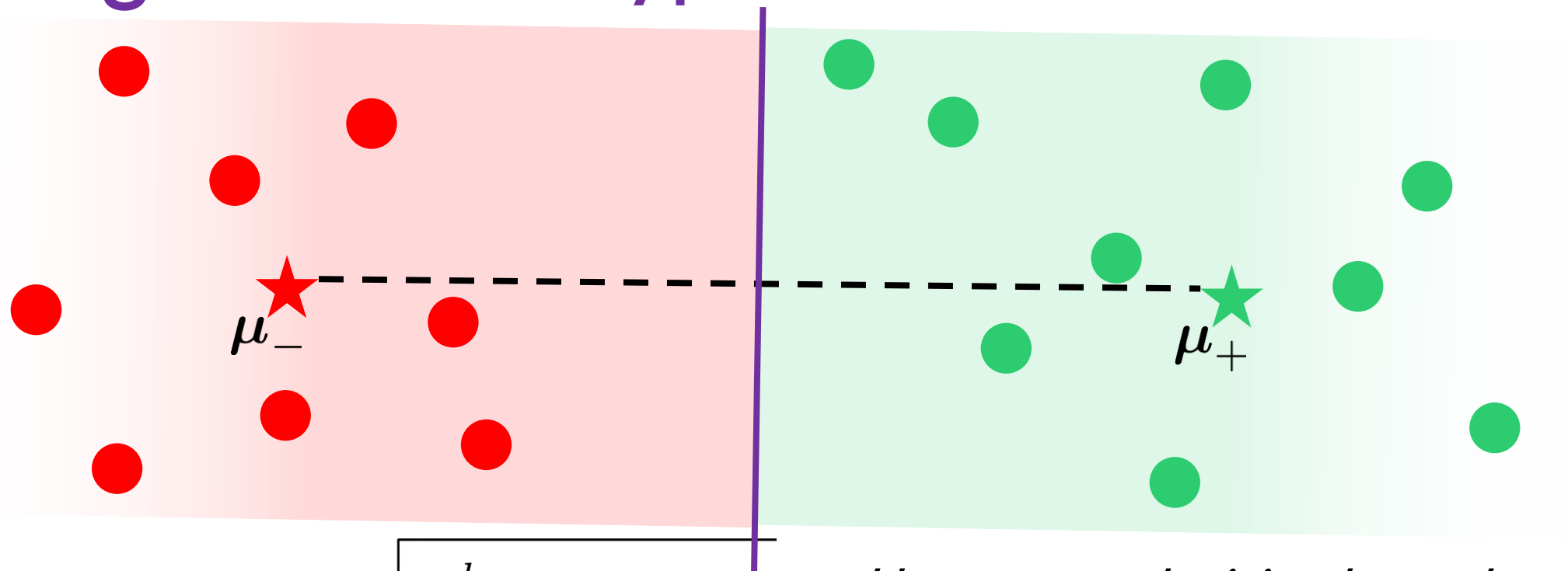
$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Linear Decision Boundary

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

*How to get decision boundary
for arbitrary metrics?*

Learning with Prototypes



$$d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2 = \sqrt{\sum_{k=1}^d (\mathbf{x}_k - \boldsymbol{\mu}_k)^2}$$

Linear Decision Boundary

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

How to get decision boundary for arbitrary metrics?

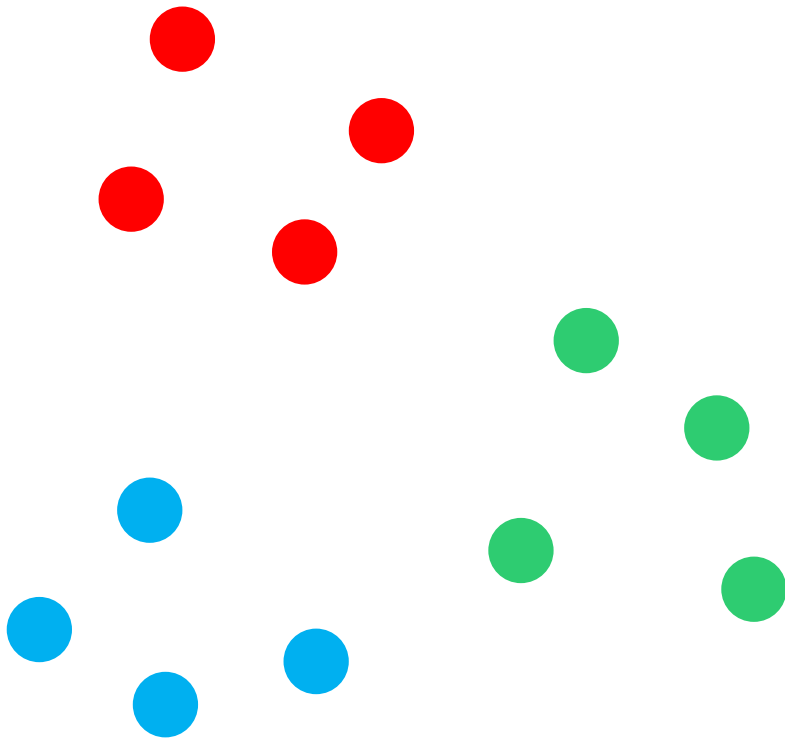
Solve for $d(x, \mu_+) \geq d(x, \mu_-)$

You will need a bit of linear algebra for this!

Multi-classification Loss Functions

One-vs-All (OVA)

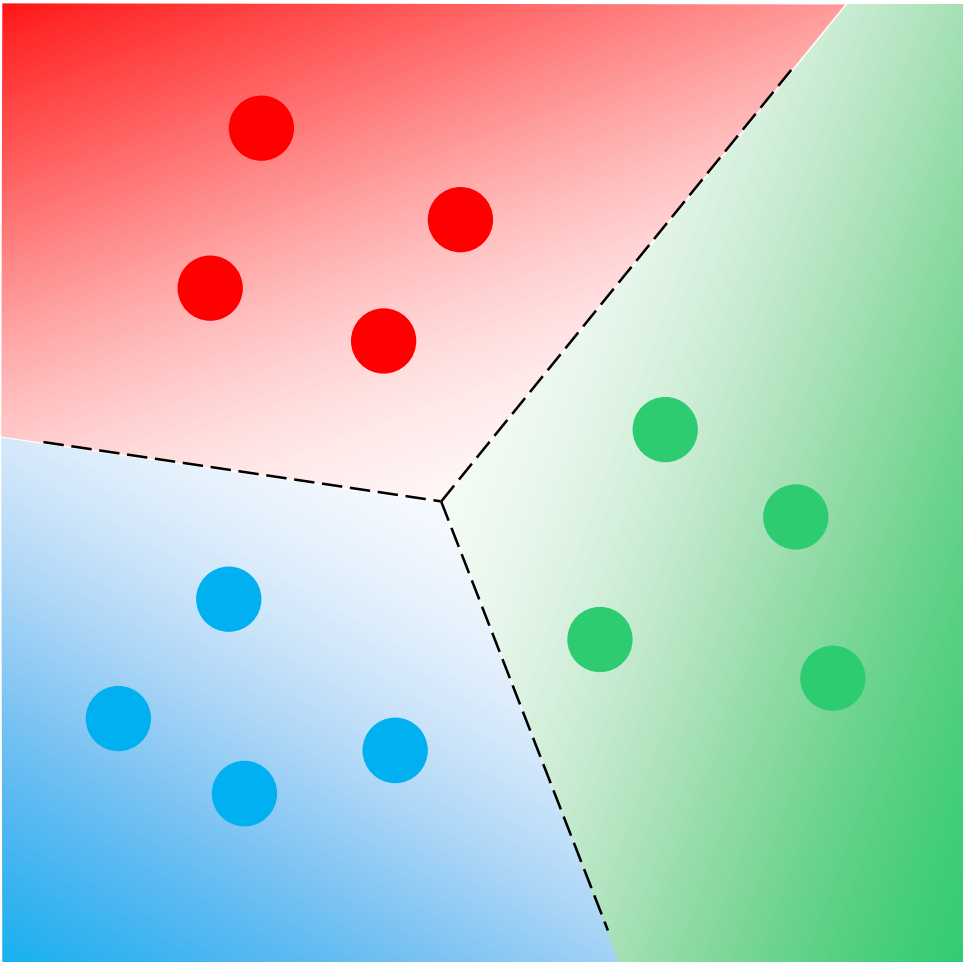
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

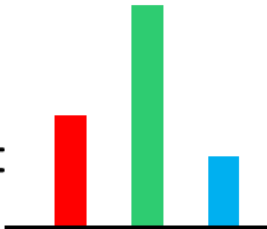
One-vs-All (OVA)

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



Multi-classification using MLE

- $K > 2$ classes – need more detailed parameters
- For each point, its label profile is a vector

$$\boldsymbol{\eta}(\mathbf{x}) =$$


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

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

- Likelihood function is multinomial instead of binomial

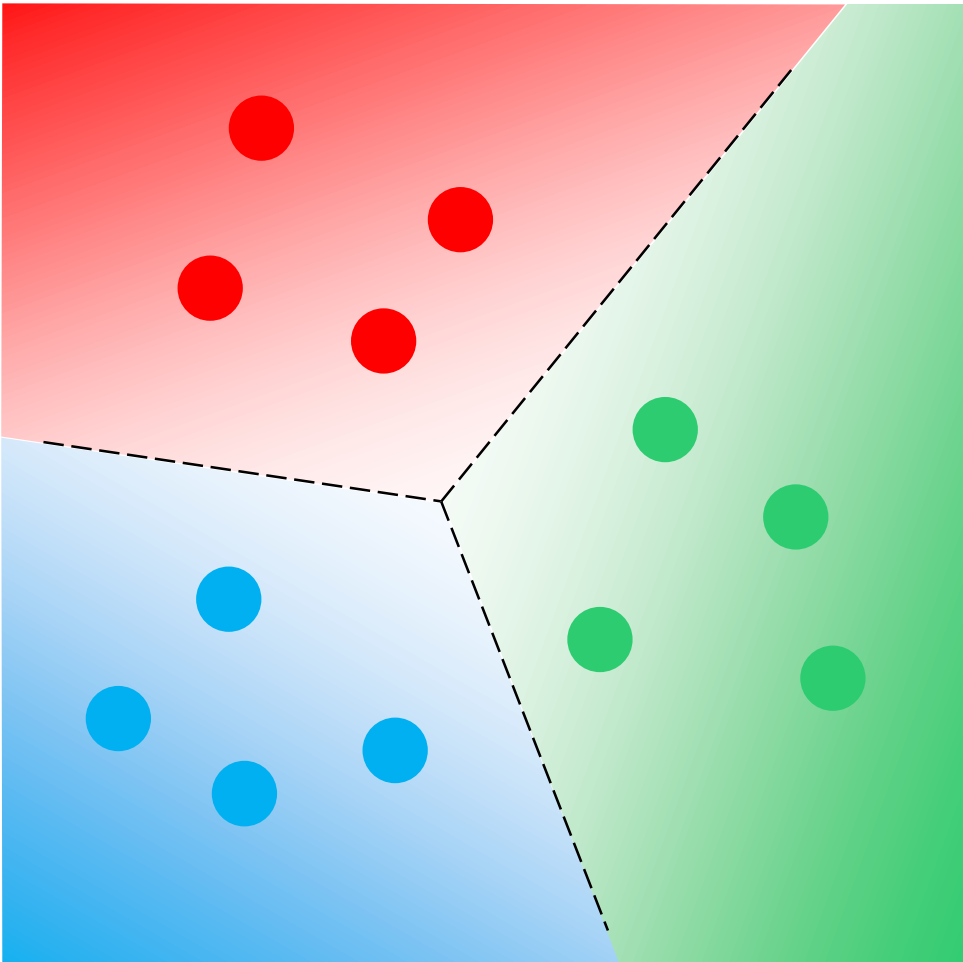
$$\mathbb{P} [\mathbf{y} \mid \mathbf{X}, \mathbf{w}] = \prod_{i=1}^n \hat{\eta}_{y^i}^i(\mathbf{x}) \quad \hat{\eta}_k^i(\mathbf{x}) = \frac{\exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)}{\sum_{l=1}^K \exp(\langle \mathbf{w}^l, \mathbf{x}^i \rangle)}$$

Softmax Regression

Multi-classification Loss Functions

One-vs-All (OVA)

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

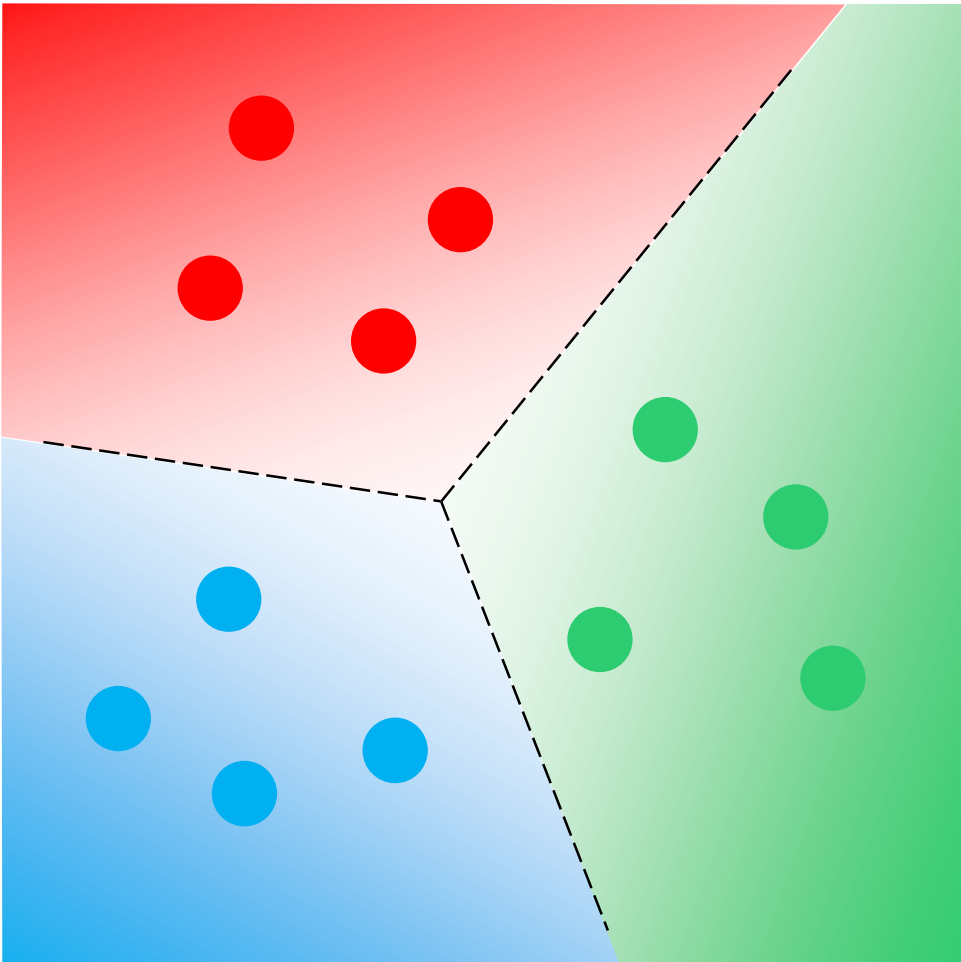


Multi-classification Loss Functions

One-vs-All (OVA)

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

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



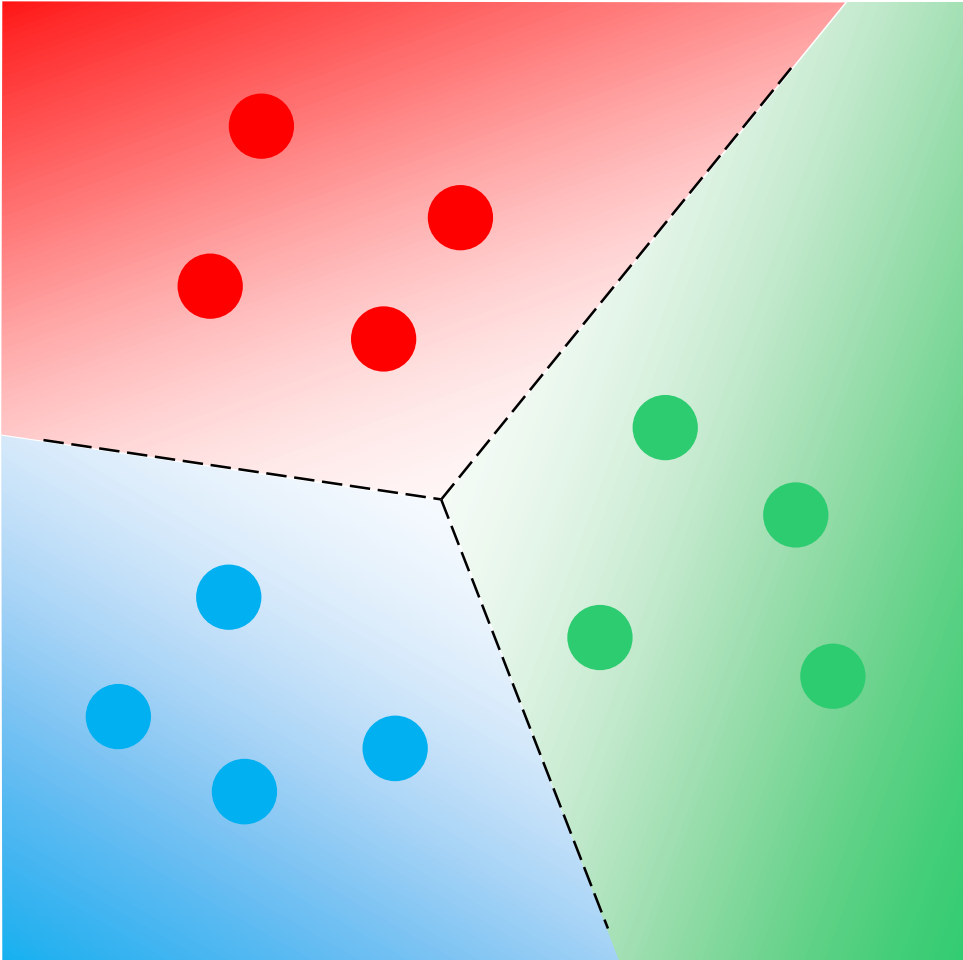
Multi-classification Loss Functions

One-vs-All (OVA)

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

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

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



Multi-classification Loss Functions

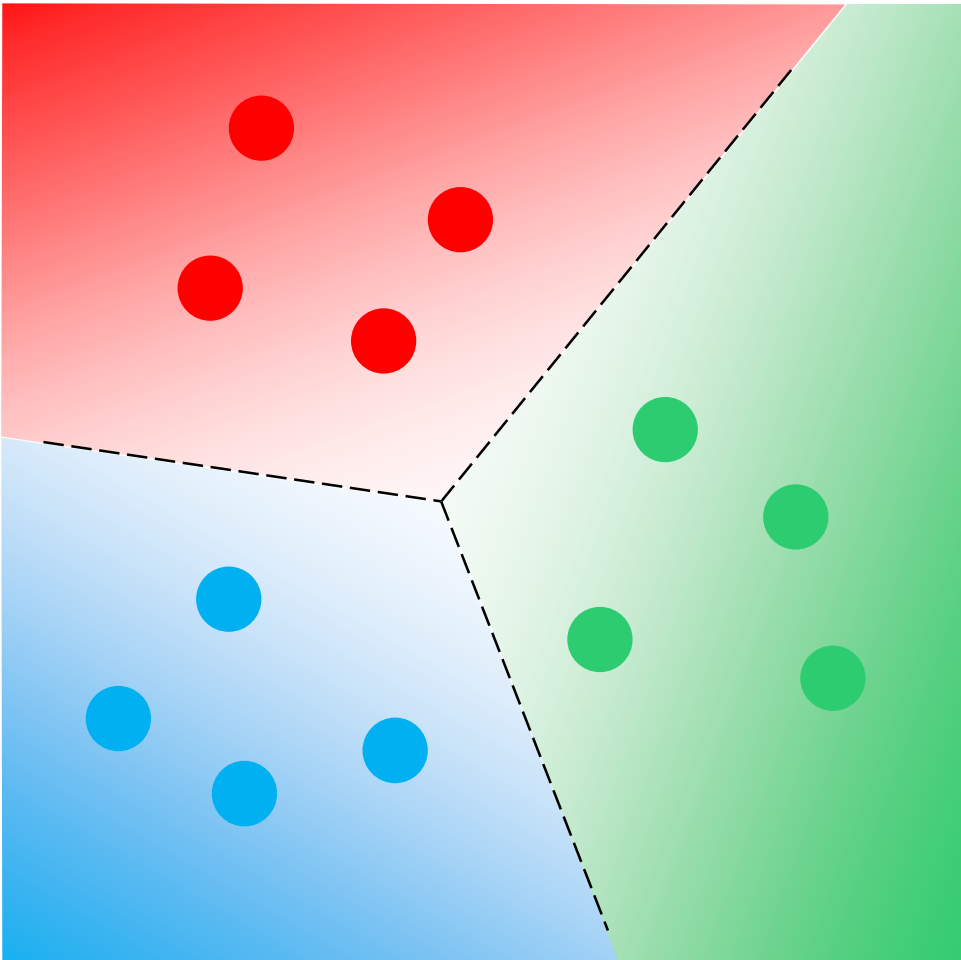
One-vs-All (OVA)

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

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

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$



Multi-classification Loss Functions

One-vs-All (OVA)

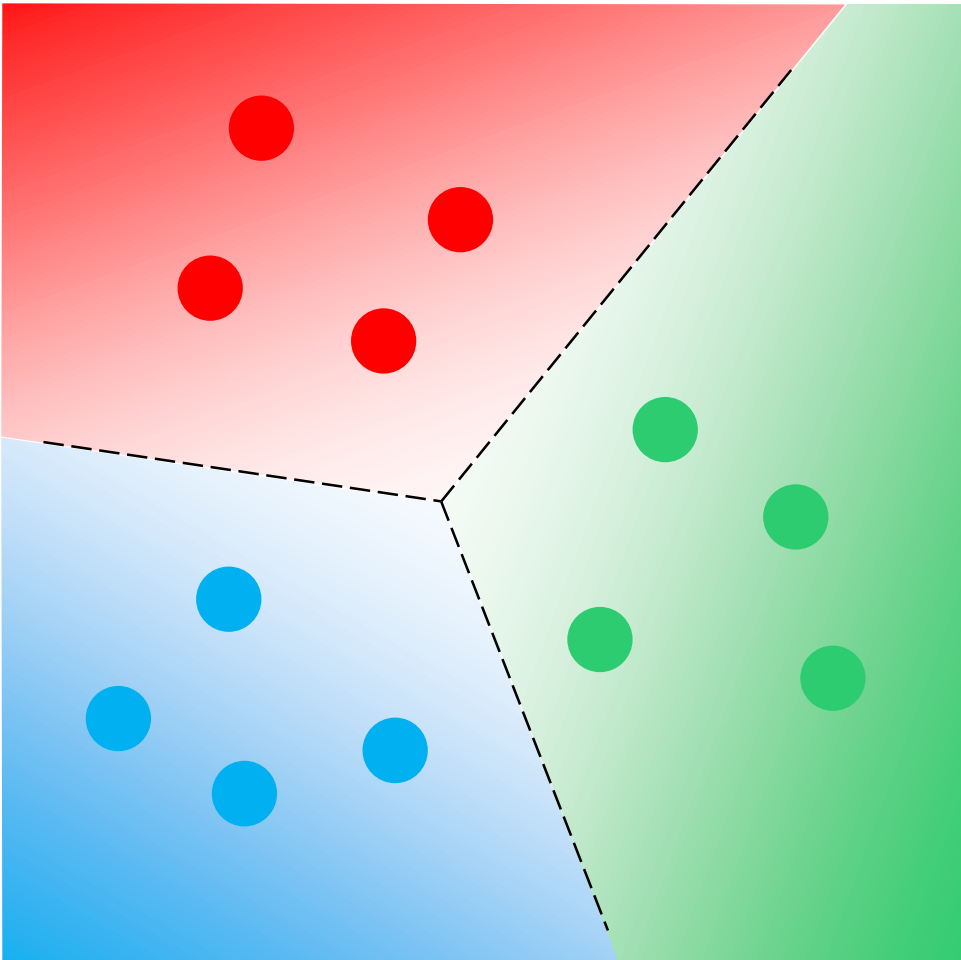
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$

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

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$

$$\ell_{\text{sm}}(y, \{\eta_j\}) = -\log \frac{\exp(\eta_y)}{\sum_{k=1}^K \exp(\eta_k)}$$



Multi-classification Loss Functions

One-vs-All (OVA)

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

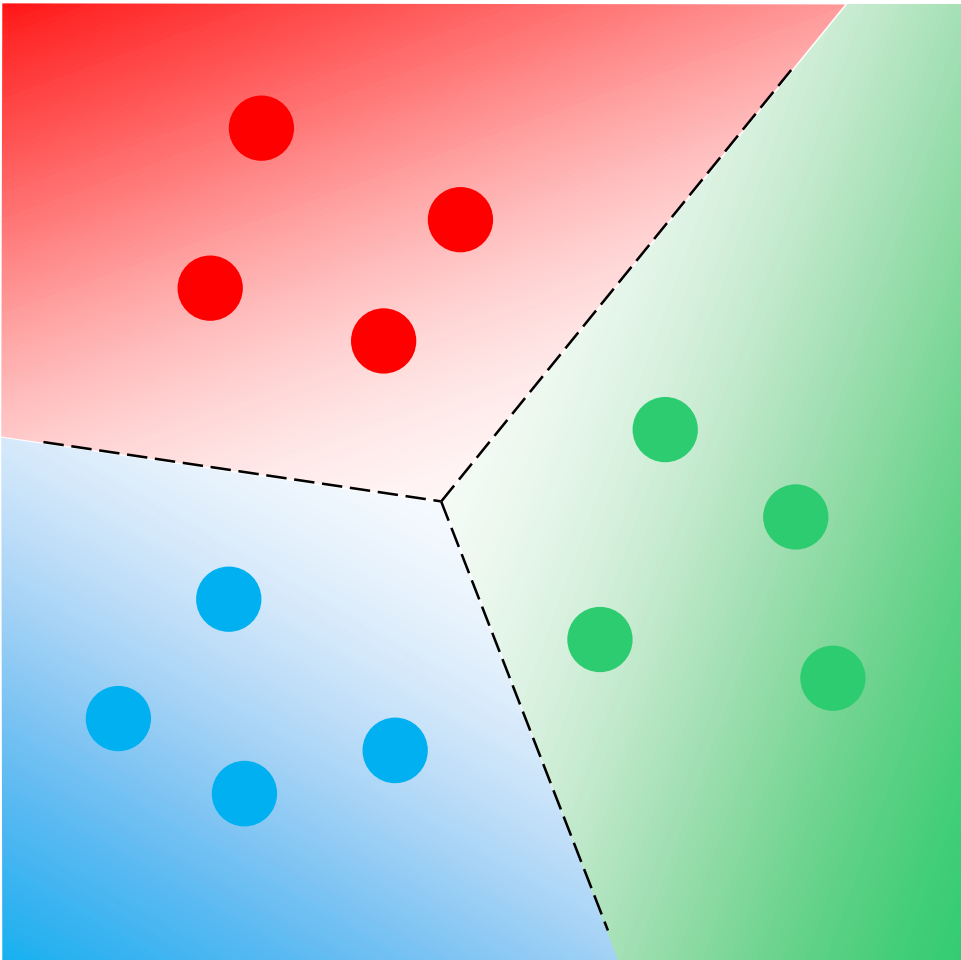
$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^K]$$

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$

$$\ell_{\text{sm}}(y, \{\eta_j\}) = -\log \frac{\exp(\eta_y)}{\sum_{k=1}^K \exp(\eta_k)}$$

Softmax loss
function

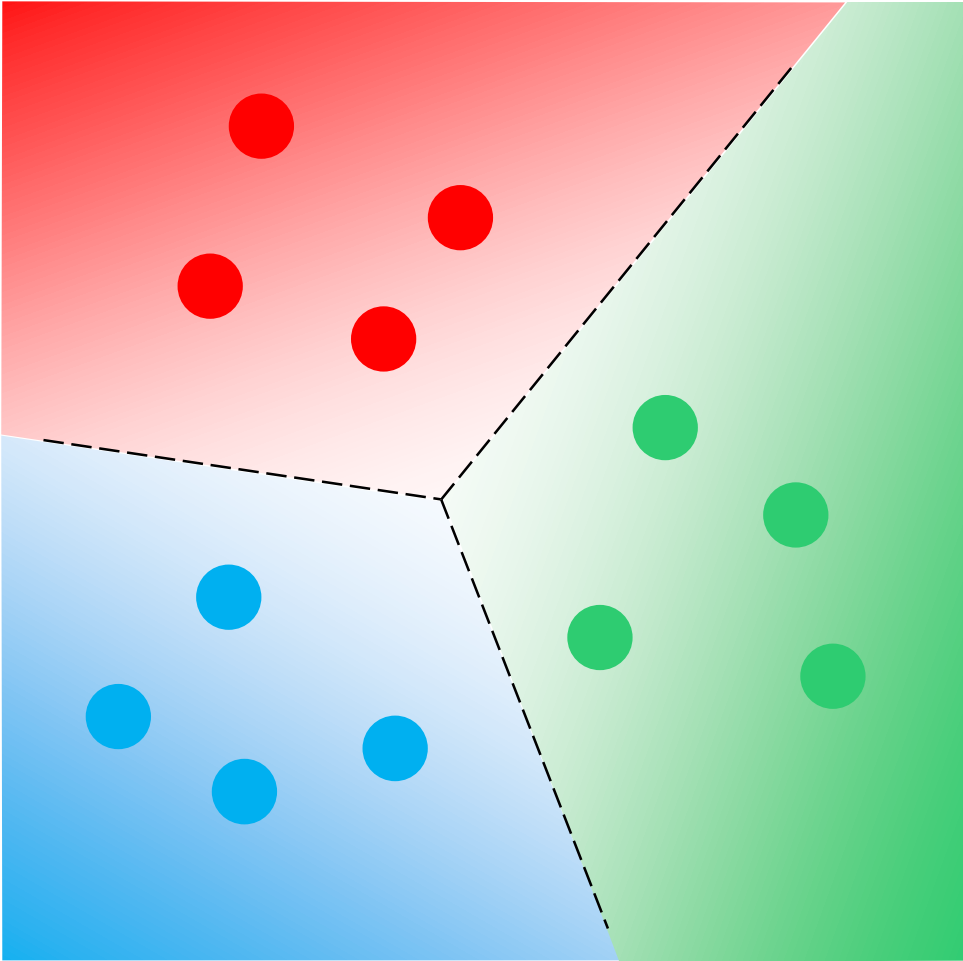


Multi-classification Loss Functions

One-vs-All (OVA)

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

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



Multi-classification Loss Functions

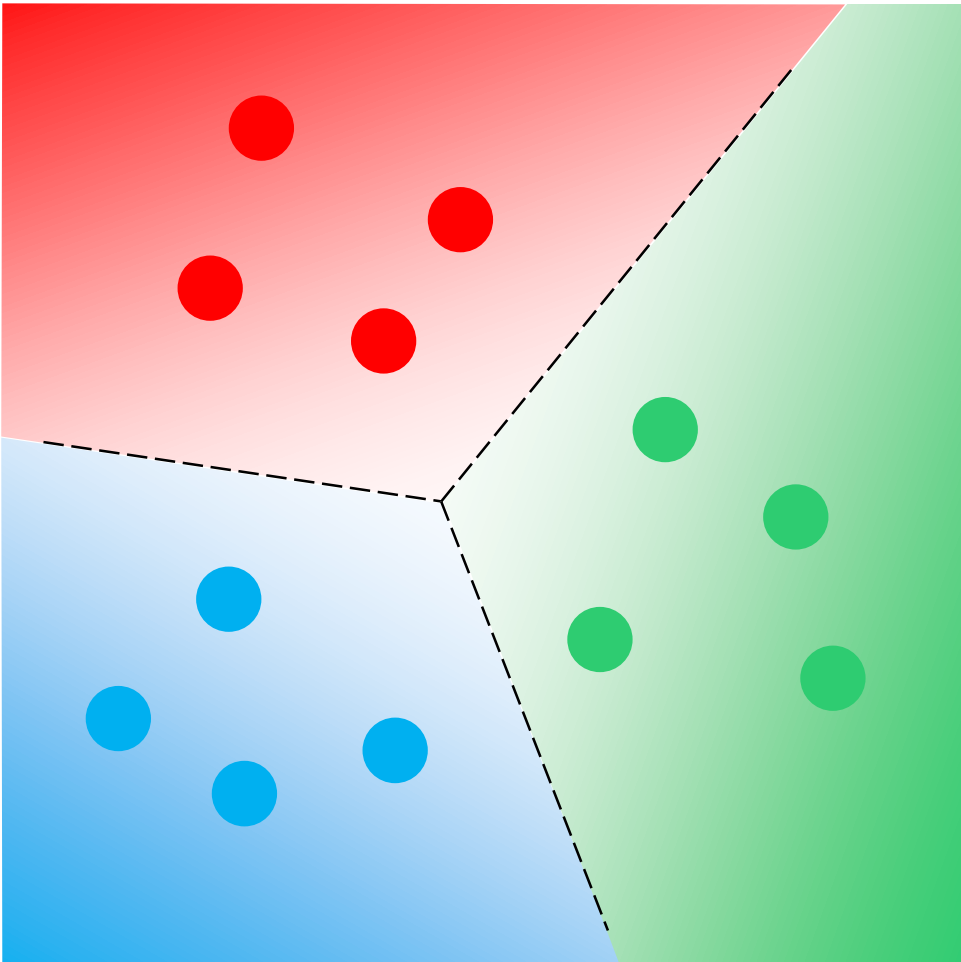
One-vs-All (OVA)

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

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

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

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



Multi-classification Loss Functions

One-vs-All (OVA)

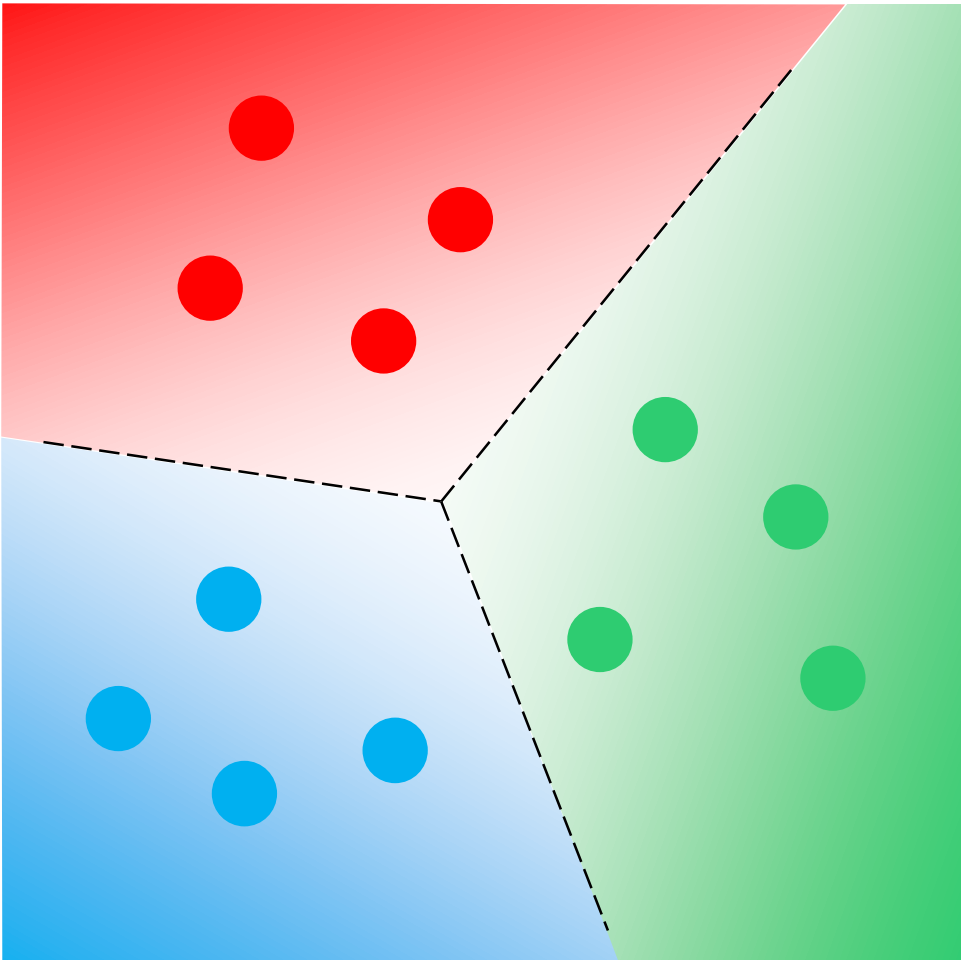
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$

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

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

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

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



Multi-classification Loss Functions

One-vs-All (OVA)

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

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

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

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

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

Slack variable



Multi-classification Loss Functions

One-vs-All (OVA)

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

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

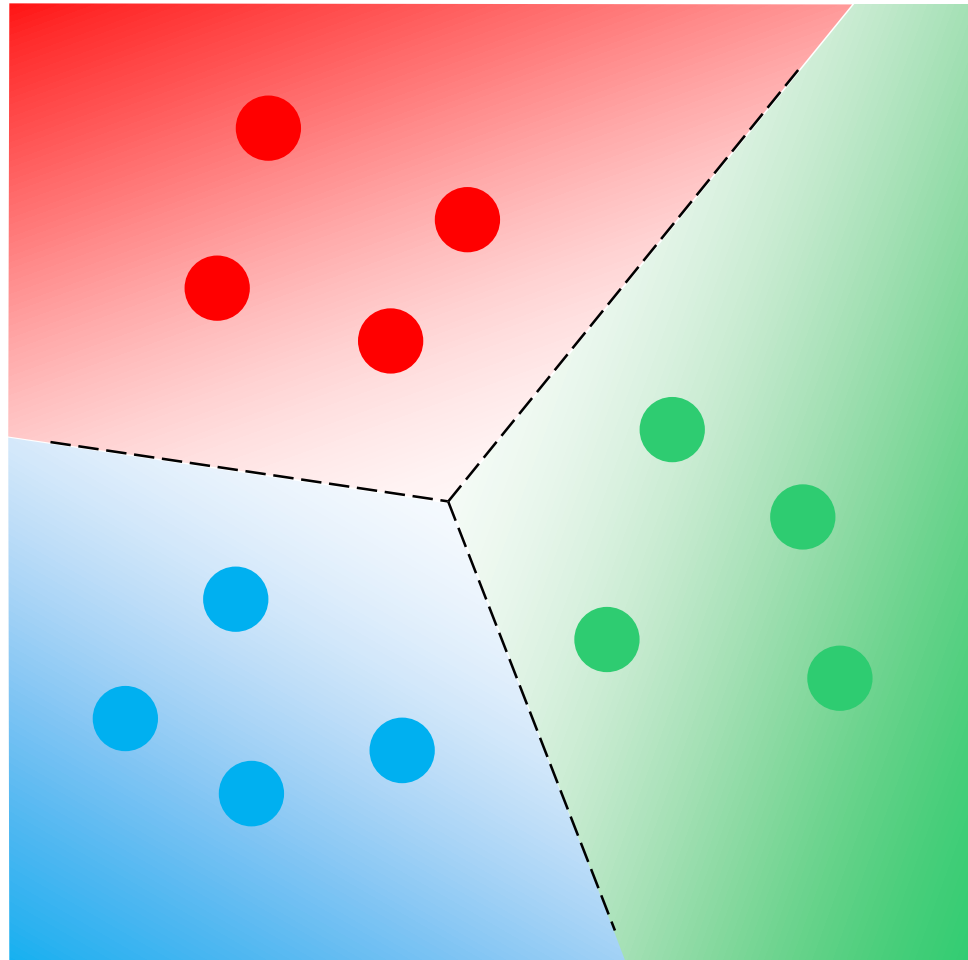
$$\hat{\mathbf{W}} = \arg \min_{\mathbf{W}} \sum_{k=1}^K \|\mathbf{w}^k\|_2^2$$

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

$$\xi_i \geq 0$$

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

Slack variable



Multi-classification Loss Functions

One-vs-All (OVA)

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

Slack variable

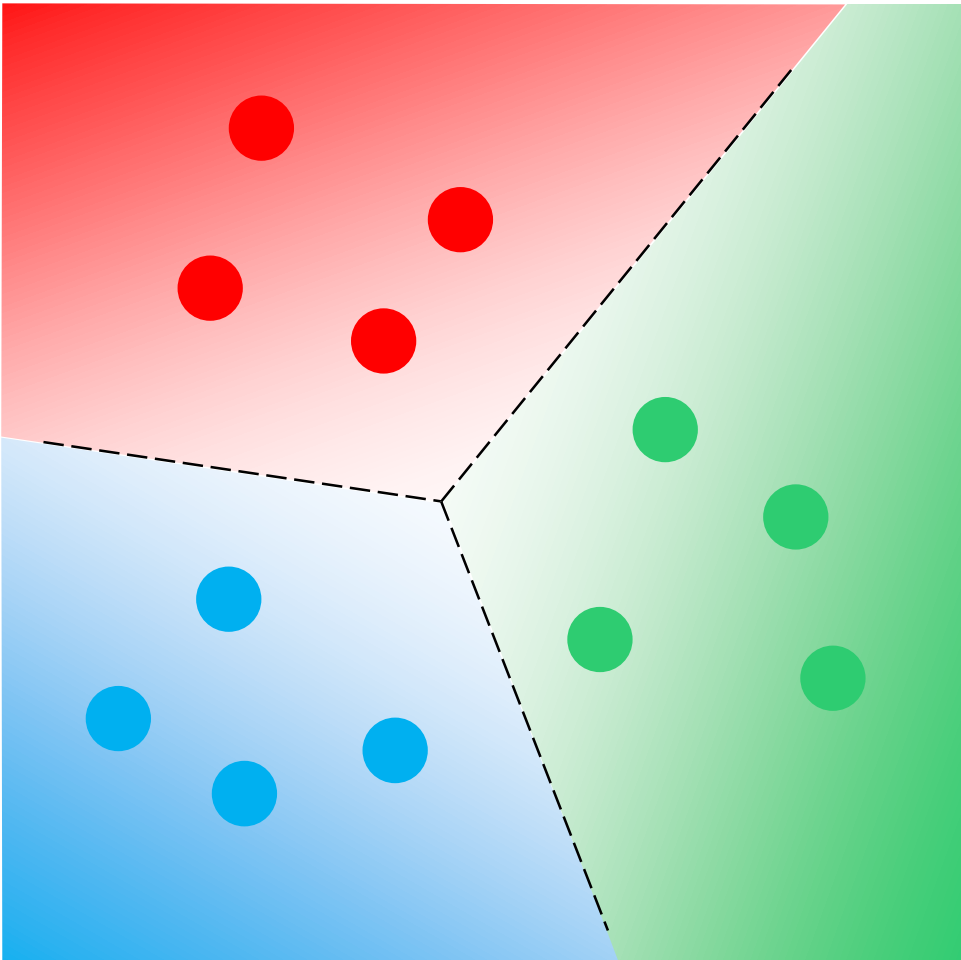
$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^K]$$

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

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

$$\xi_i \geq 0$$

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



Multi-classification Loss Functions

One-vs-All (OVA)

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

Slack variable

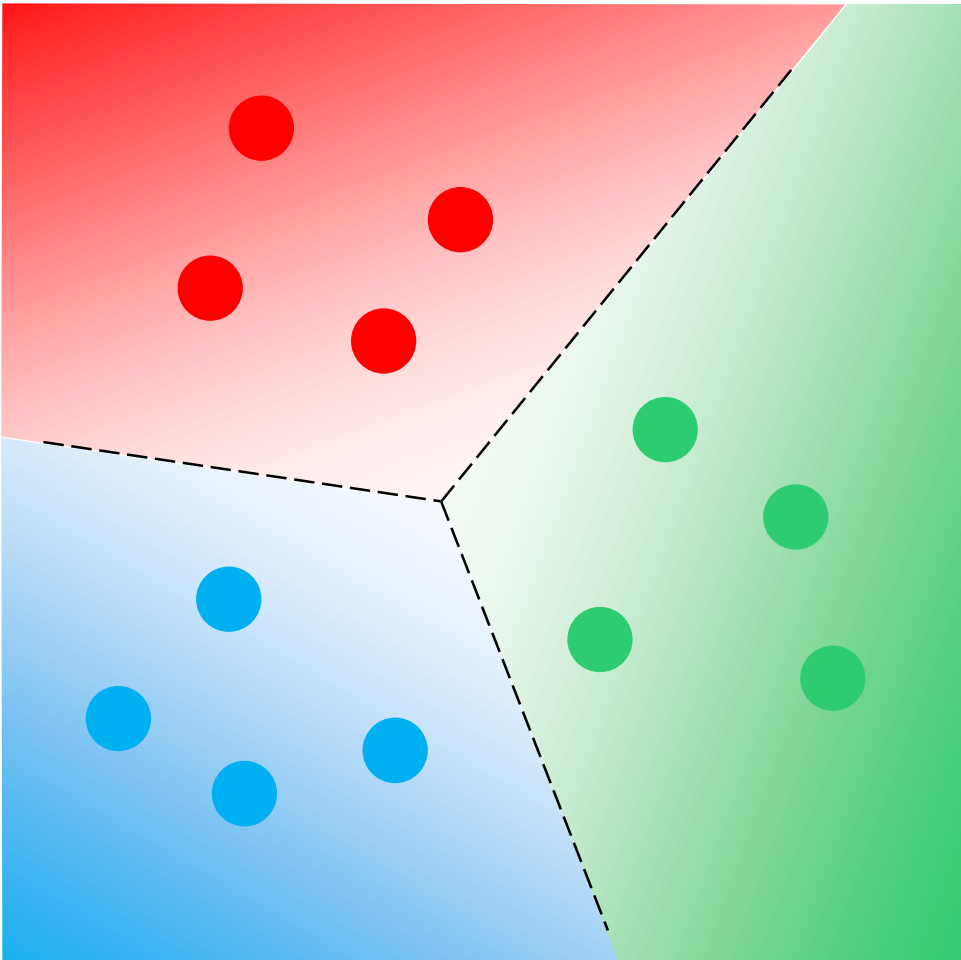
$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^K]$$

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

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

$$\xi_i \geq 0$$

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



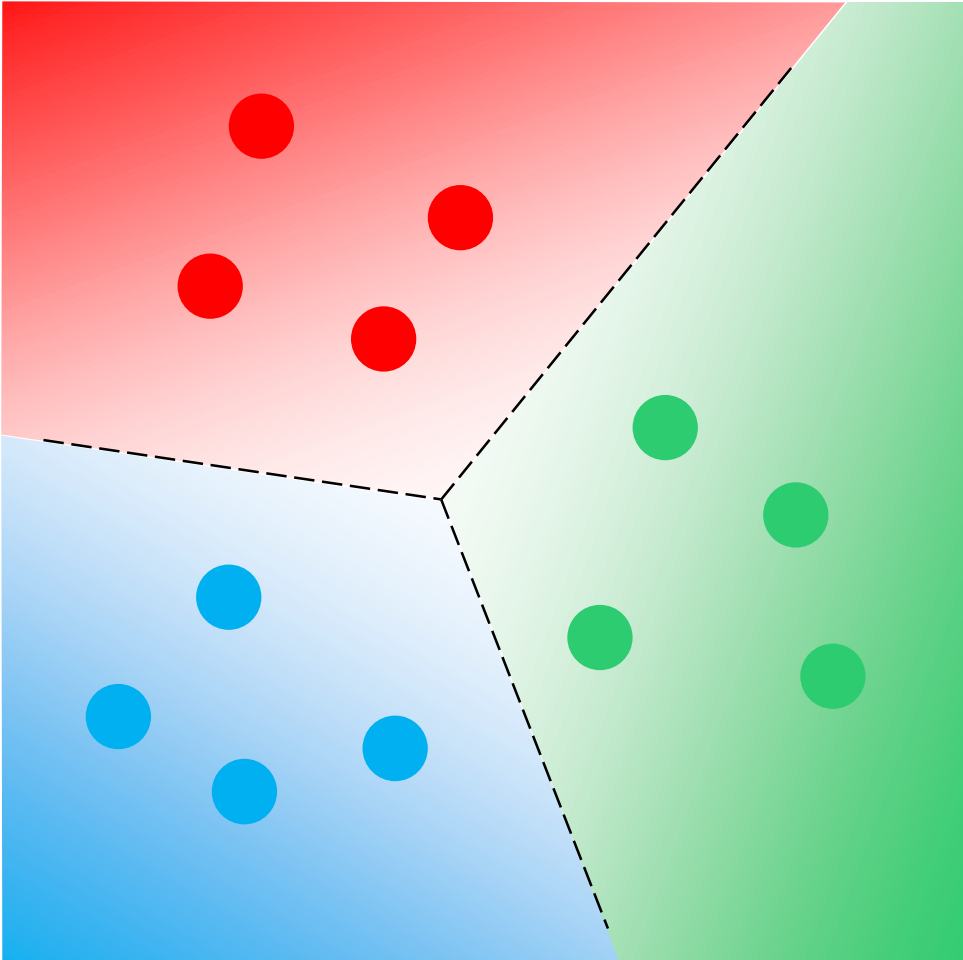
Multi-classification Loss Functions

One-vs-All (OVA)

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

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

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



Multi-classification Loss Functions

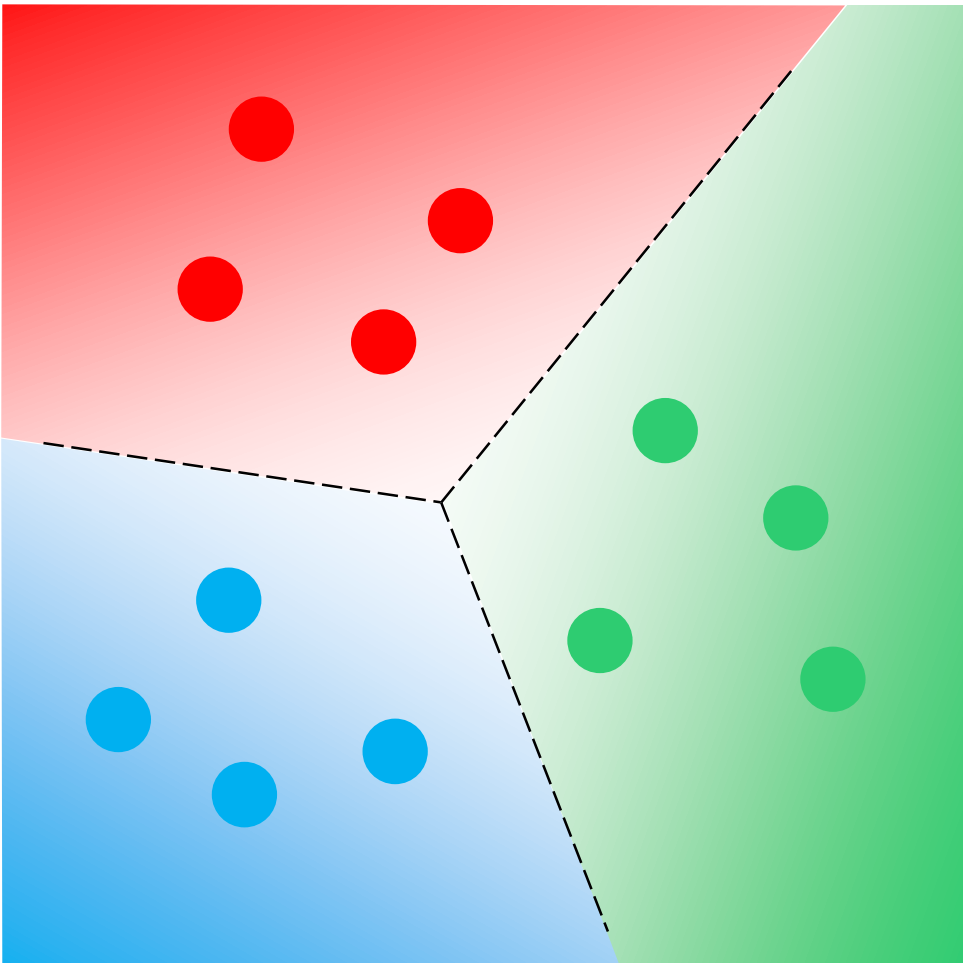
One-vs-All (OVA)

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

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

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$



Multi-classification Loss Functions

One-vs-All (OVA)

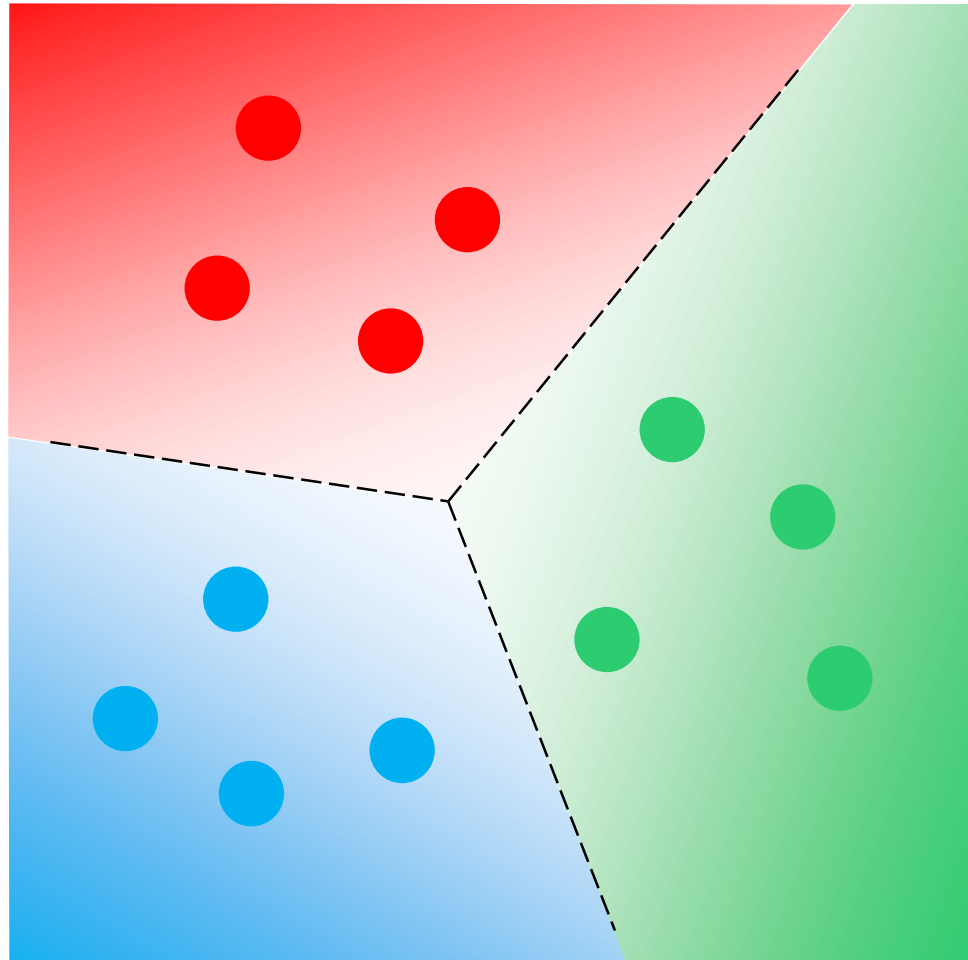
$$\hat{y}^i = \arg \max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$

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

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$

$$\ell_{\text{cs}}(y, \{\eta_j\}) = [1 + \max_{k \neq y} \eta_k - \eta_y]_+$$



Multi-classification Loss Functions

One-vs-All (OVA)

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

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

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$

$$\ell_{\text{cs}}(y, \{\eta_j\}) = [1 + \max_{k \neq y} \eta_k - \eta_y]_+$$

Assignment
problem

Multi-classification Loss Functions

One-vs-All (OVA)

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

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

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

$$\langle \mathbf{W}, \mathbf{x} \rangle = \boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$$

$$\ell_{\text{cs}}(y, \{\eta_j\}) = [1 + \max_{k \neq y} \eta_k - \eta_y]_+$$

Assignment
problem

Crammer-Singer
loss function

Please give your Feedback

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