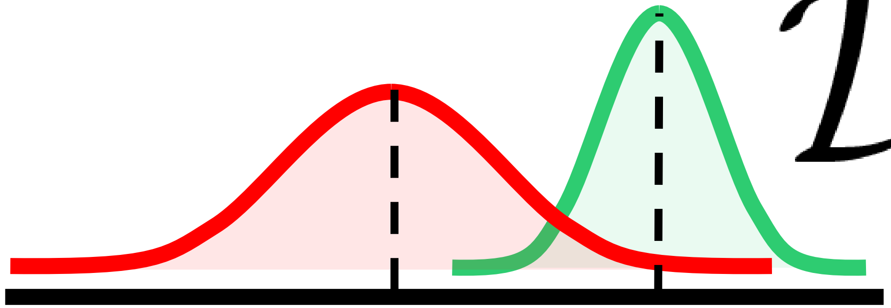
Function Approximation Methods-I

CS771: Introduction to Machine Learning

Purushottam Kar

Recap



$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Longleftarrow} \mathbf{Z}$$

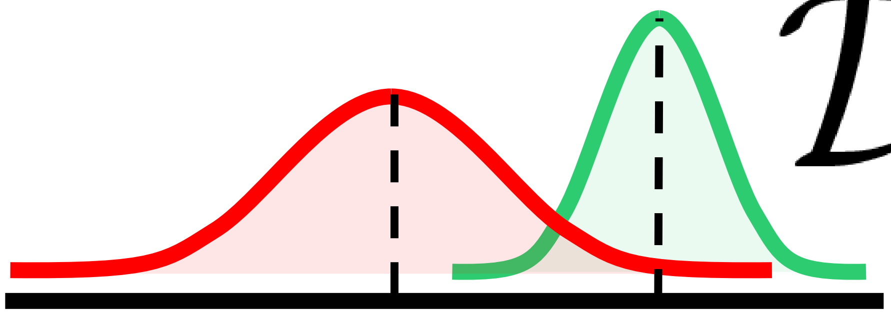
$\mathbb{P}[\boldsymbol{\theta}]$	$\mathbb{P}[\mathbf{z} \mid \boldsymbol{\theta}]$	$\mathbb{P}[\boldsymbol{\theta} \mid \mathbf{Z}]$	$\mathbb{P}[\mathbf{z} \mid \mathbf{Z}]$
Prior	Likelihood	Posterior	Predictive Posterior

$$\mathbb{P}[\mathbf{z} \mid \mathbf{Z}] = \int_{\boldsymbol{\theta}} \mathbb{P}[\mathbf{z} \mid \boldsymbol{\theta}]\, \mathbb{P}[\boldsymbol{\theta} \mid \mathbf{Z}]\, d\boldsymbol{\theta}$$

MAP, MLE

“Challenging” integral

Recap



$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Leftarrow} \mathbf{Z}$$

$$\mathbb{P}\left[\boldsymbol{\theta}\right] \qquad \mathbb{P}\left[\mathbf{z} \,\middle|\, \boldsymbol{\theta}\right] \qquad \mathbb{P}\left[\boldsymbol{\theta} \,\middle|\, \mathbf{Z}\right] \qquad \mathbb{P}\left[\mathbf{z} \,\middle|\, \mathbf{Z}\right]$$

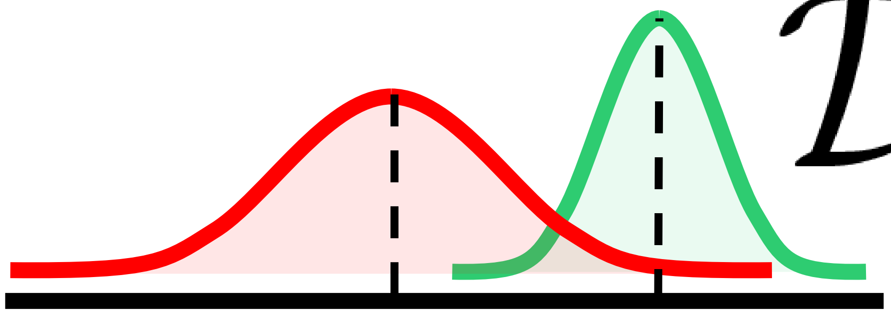
Prior | Likelihood | Posterior | Predictive Posterior

$$\mathrm{Beta}(p \,|\, \alpha, \beta) \qquad \mathrm{Beta}(p \,|\, \alpha + n_H, \beta + n_T)$$

$$\mathrm{Bernoulli}(y \,|\, p) \qquad \mathrm{Bernoulli}\left(y \,\middle|\, \frac{\alpha + n_H}{\alpha + \beta + n}\right)$$

Bias Estimation

Recap



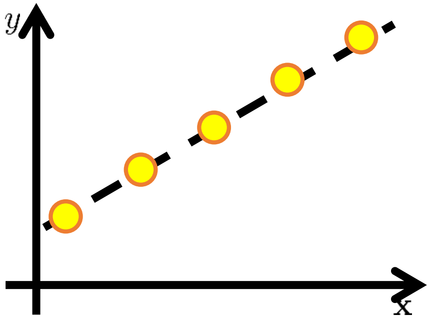
$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Leftarrow} \mathbf{Z}$$

$\mathbb{P}\left[\boldsymbol{\theta}\right]$	$\mathbb{P}\left[\mathbf{z} \mid \boldsymbol{\theta}\right]$	$\mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{Z}\right]$	$\mathbb{P}\left[\mathbf{z} \mid \mathbf{Z}\right]$
Prior	Likelihood	Posterior	Predictive Posterior

$$\mathcal{N}(\mathbf{w} \mid \mathbf{0}, \rho) \propto \exp\left(-\|\mathbf{w}\|_2^2/2\rho^2\right)$$

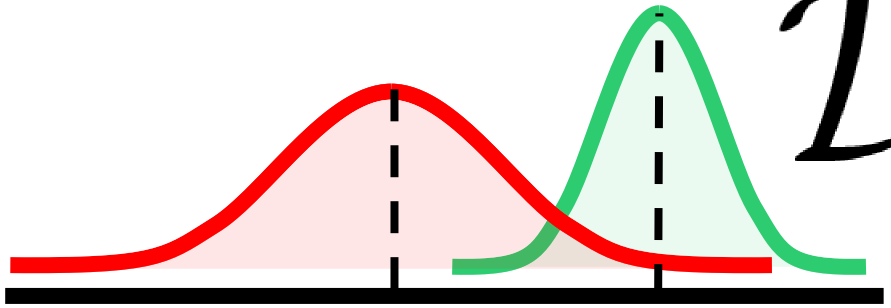
$$\mathrm{Lap}(\mathbf{w} \mid \mathbf{0}, \gamma) \propto \exp\left(-\gamma \cdot \|\mathbf{w}\|_1\right)$$

Sparsity inducing prior



Linear Regre

Recap



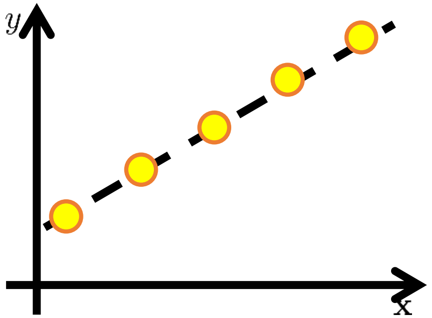
$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Longleftarrow} \mathbf{Z}$$

$\mathbb{P}[\boldsymbol{\theta}]$	$\mathbb{P}[\mathbf{z} \mid \boldsymbol{\theta}]$	$\mathbb{P}[\boldsymbol{\theta} \mid \mathbf{Z}]$	$\mathbb{P}[\mathbf{z} \mid \mathbf{Z}]$
Prior	Likelihood	Posterior	Predictive Posterior

$$\mathcal{N}(\mathbf{w} \mid \mathbf{0}, \rho) \propto \exp\left(-\|\mathbf{w}\|_2^2 / 2\rho^2\right)$$

$$\mathcal{N}(\mathbf{w} \mid \boldsymbol{\nu}, \Lambda)$$

$$\mathbb{P}[y \mid \mathbf{x}, \mathbf{X}, \mathbf{y}]$$

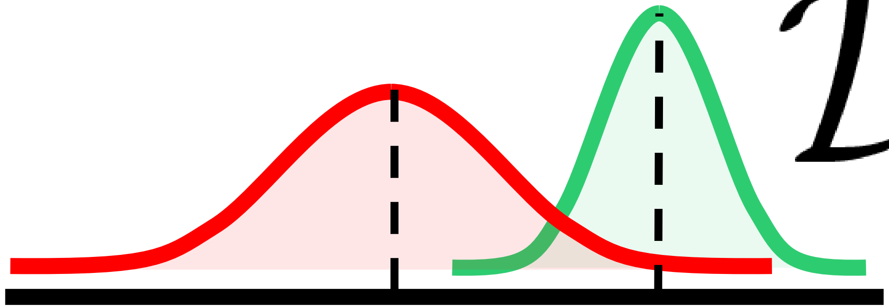


$$\mathcal{N}(y \mid \langle \mathbf{w}, \mathbf{x} \rangle, \sigma)$$

$$\mathcal{N}\left(y \mid \langle \boldsymbol{\nu}, \mathbf{x} \rangle, \sigma^2 + \mathbf{x}^\top \Lambda \mathbf{x}\right)$$

Linear Regression

Recap



$$\mathcal{D}_{\boldsymbol{\theta}^*} \xleftarrow{?} \mathbf{Z}$$

$$\mathbb{P}\left[\boldsymbol{\theta}\right] \quad \mathbb{P}\left[\mathbf{z} \mid \boldsymbol{\theta}\right] \quad \mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{Z}\right] \quad \mathbb{P}\left[\mathbf{z} \mid \mathbf{Z}\right]$$

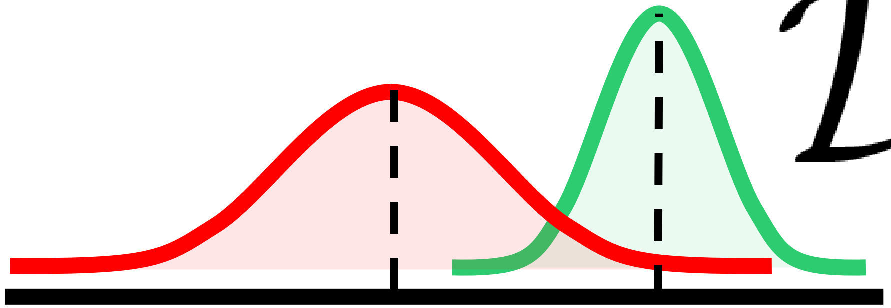
Prior Likelihood Posterior Predictive Posterior

$$\mathcal{N}(\mathbf{w} \mid \mathbf{0}, \rho) \propto \exp\left(-\|\mathbf{w}\|_2^2 / 2\rho^2\right)$$

Can have sparsity inducing prior also!!

Linear Classif

Recap

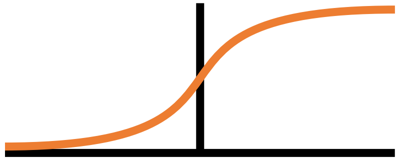


$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Longleftarrow} \mathbf{Z}$$

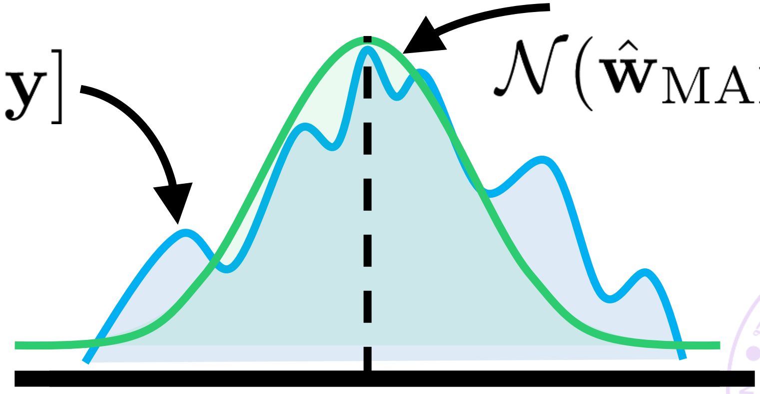
$\mathbb{P}\left[\boldsymbol{\theta}\right]$	$\mathbb{P}\left[\mathbf{z} \mid \boldsymbol{\theta}\right]$	$\mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{Z}\right]$	$\mathbb{P}\left[\mathbf{z} \mid \mathbf{Z}\right]$
Prior	Likelihood	Posterior	Predictive Posterior

$$\mathcal{N}(\mathbf{w} \mid \mathbf{0}, \rho) \propto \exp\left(-\|\mathbf{w}\|_2^2 / 2\rho^2\right)$$

$$\text{Rademacher}(y \mid \sigma\left(\langle \mathbf{w}, \mathbf{x}^i \rangle\right))$$

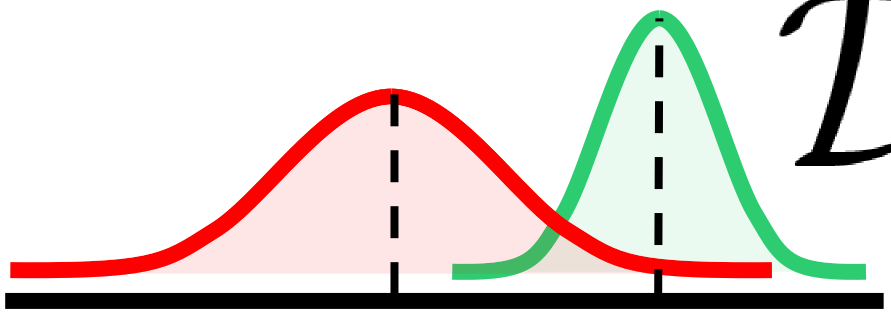


$$\text{Bernoulli}(y \mid \sigma\left(\langle \mathbf{w}, \mathbf{x}^i \rangle\right))$$

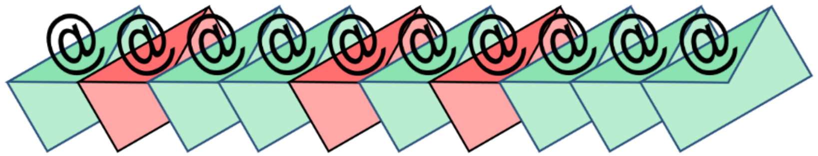


Linear Classification

Recap



$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Leftarrow} \mathbf{Z}$$

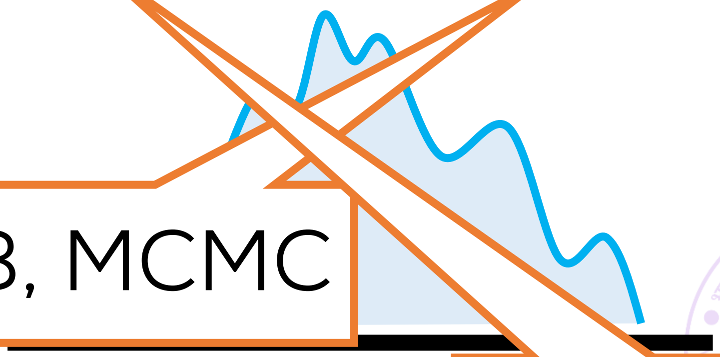


$$\mathbb{P}\left[\boldsymbol{\theta}\right] \quad \mathbb{P}\left[\mathbf{z} \mid \boldsymbol{\theta}\right] \quad \mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{Z}\right] \quad \mathbb{P}\left[\mathbf{z} \mid \mathbf{Z}\right]$$

Prior Likelihood Posterior Predictive Posterior

$$\mathcal{N}(\mathbf{w} \mid \mathbf{0}, \rho) \propto \exp\left(-\|\mathbf{w}\|_2^2 / 2\rho^2\right)$$

$$\text{Bernoulli}\left(y \;\middle|\; \frac{\exp(\langle \mathbf{w}^y, \mathbf{x}^i \rangle)}{\sum_{l=1}^K \exp(\langle \mathbf{w}^l, \mathbf{x}^i \rangle)}\right)$$



VB, MCMC

MAP, MLE

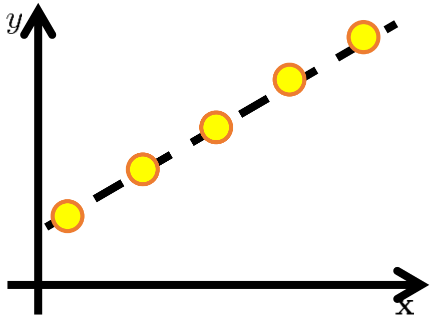
Multi-Classification

Function Approximation

In other words, here comes the Math …

Learning through Optimization

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



Learning through Optimization

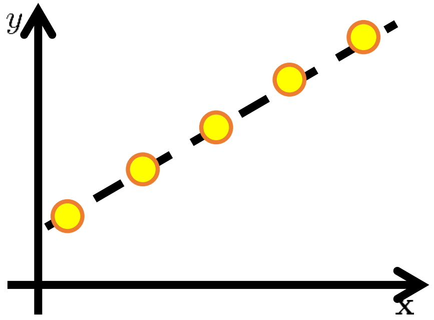
$$\hat{\mathbf{w}}_{\text{MLE}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

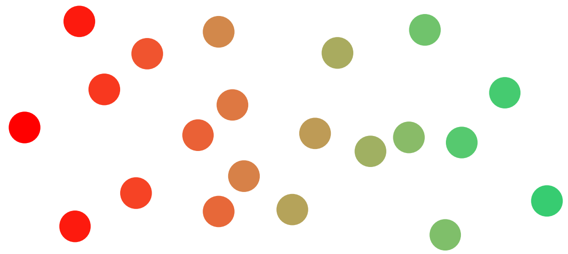
$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{\sigma^2}{\rho^2} \|\mathbf{w}\|_2^2$$

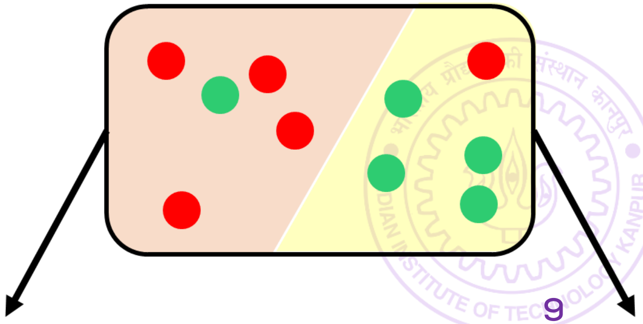
$$\hat{\mathbf{w}}_{\text{MLE}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \quad \ldots \quad\right) + \lambda \|\mathbf{w}\|_1$$

All Linear Functions??





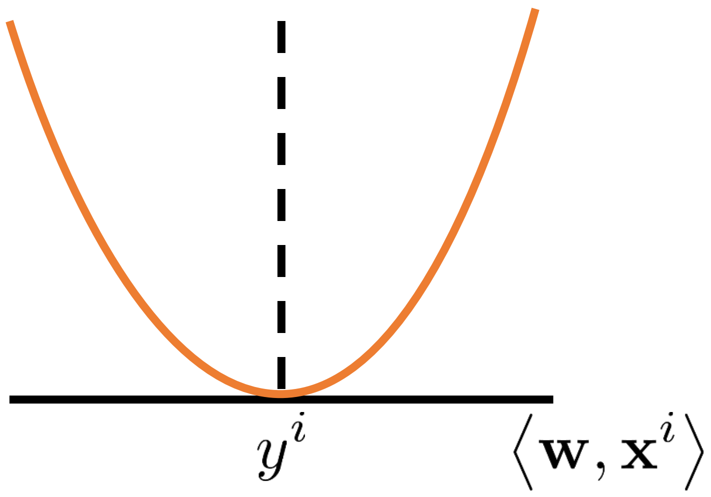


Fit and loss

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

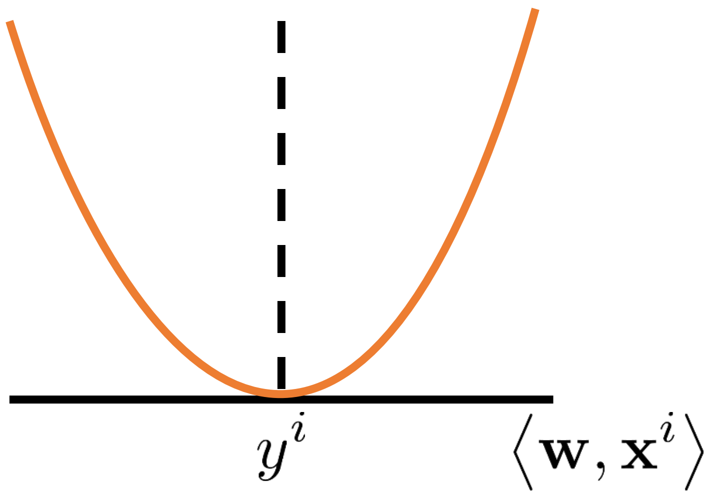
Fit and loss

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



Fit and loss

$$\ell_{\mathrm{sq}}(y, \hat{y}) = (y - \hat{y})^2$$

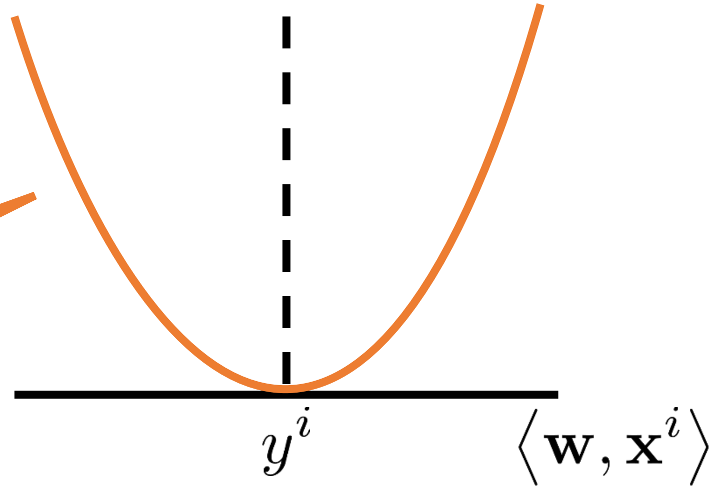


Fit and loss

$$\ell_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2$$

Popular for regression

Squared loss

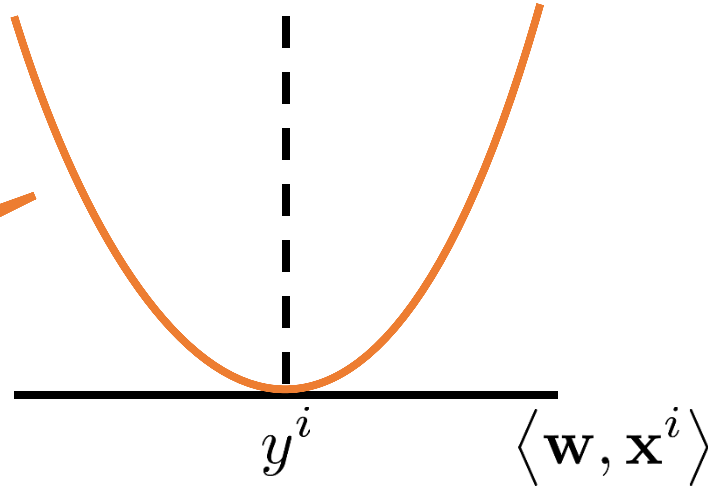


Fit and loss

$$\ell_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2$$

Popular for regression

Squared loss



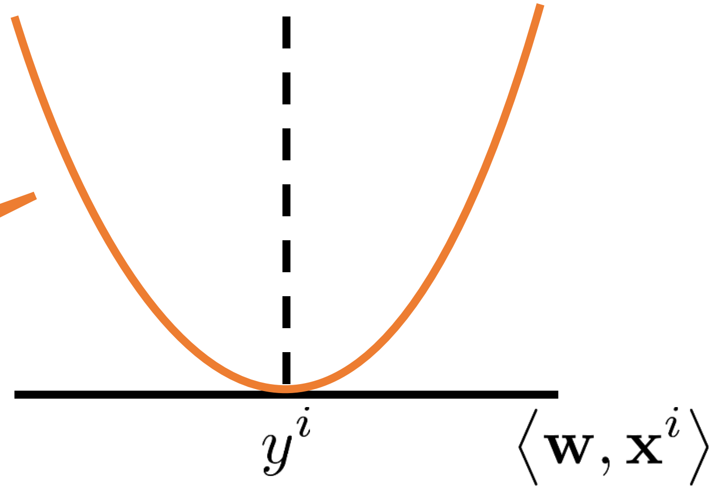
$$\hat{\mathbf{w}}_{\text{MLE}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

Fit and loss

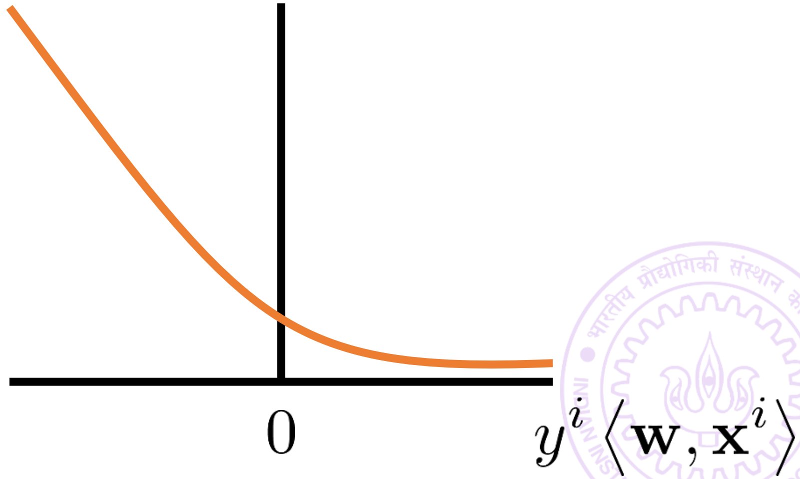
$$\ell_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2$$

Popular for regression

Squared loss



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

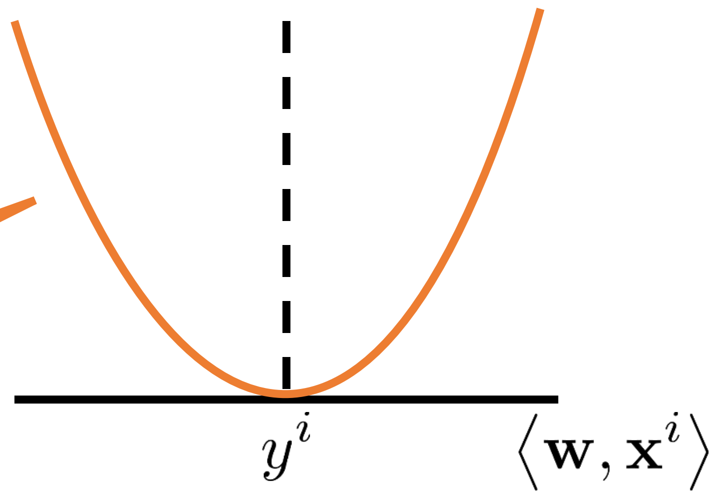


Fit and loss

$$\ell_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2$$

Popular for regression

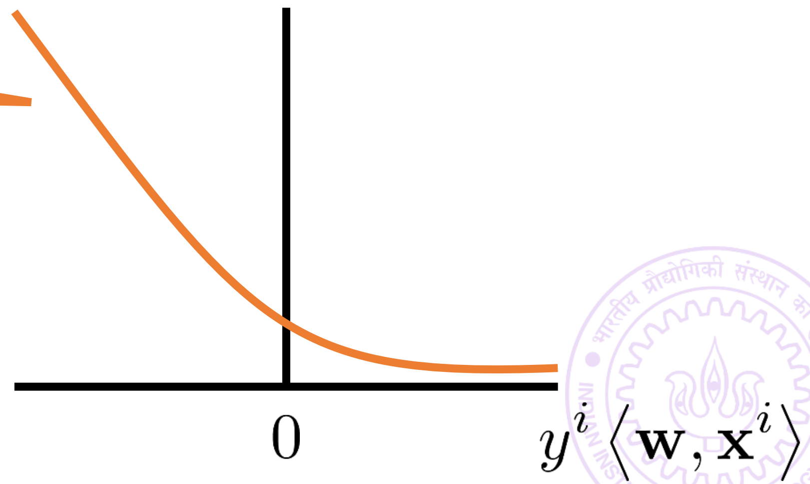
Squared loss



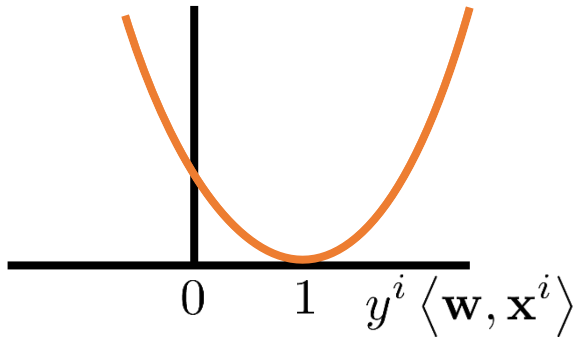
$$\ell_{\log}(y, \hat{y}) = \log(1 + \exp(-y \cdot \hat{y}))$$

Popular for classification

Logistic loss

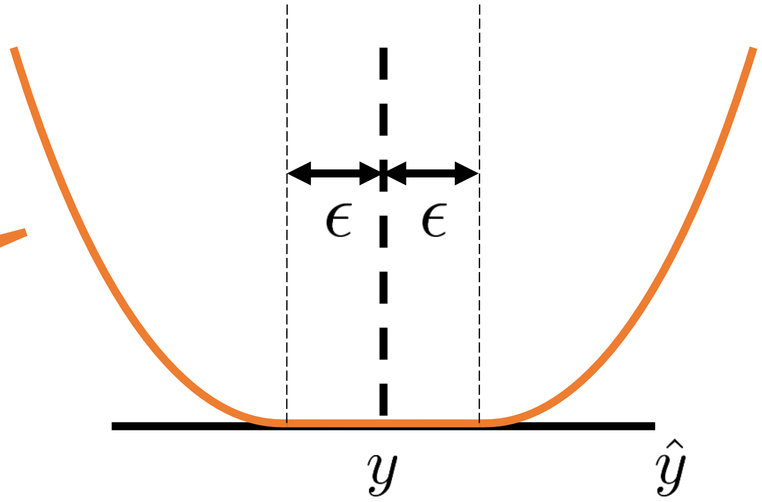


Use loss functions Interchangeably?



Other popular loss functions

$$\ell_\epsilon(y,\hat{y}) = \begin{cases} (y-\hat{y}-\epsilon)^2 & \text{if } \hat{y} < y-\epsilon \\ 0 & \text{if } \hat{y}-y \in [-\epsilon,\epsilon] \\ (y-\hat{y}+\epsilon)^2 & \text{if } \hat{y} > y+\epsilon \end{cases}$$



Popular for regression

Vapnik's ϵ-insensitive loss

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

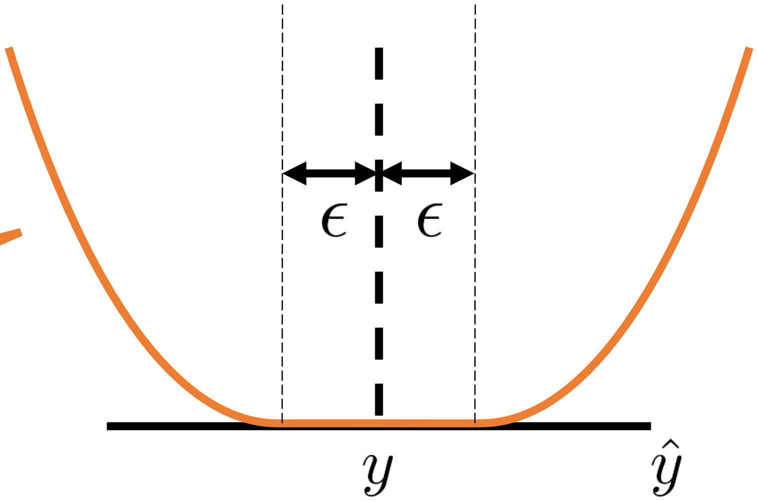
Is this an MLE?

Other popular loss functions

$$\ell_\epsilon(y, \hat{y}) = \begin{cases} (y - \hat{y} - \epsilon)^2 & \text{if } \hat{y} < y - \epsilon \\ 0 & \text{if } \hat{y} - y \in [-\epsilon, \epsilon] \\ (y - \hat{y} + \epsilon)^2 & \text{if } \hat{y} > y + \epsilon \end{cases}$$

Popular for regression

Vapnik's ϵ-insensitive loss



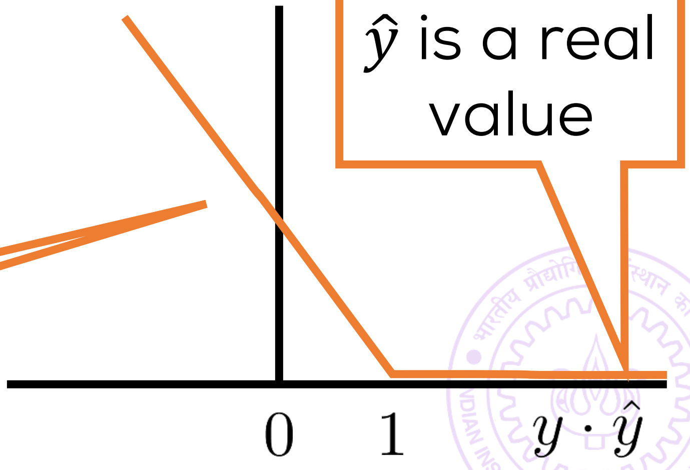
$$\ell_{\text{hinge}}(y, \hat{y}) = [1 - y \cdot \hat{y}]_+ = \begin{cases} 0 & \text{if } y \cdot \hat{y} \geq 1 \\ 1 - y \cdot \hat{y} & \text{if } y \cdot \hat{y} < 1 \end{cases}$$

Popular for classification

"Margin" loss function

Hinge loss

$\hat{y}$ is a real value



Regularization

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

Regularization

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

Regularization

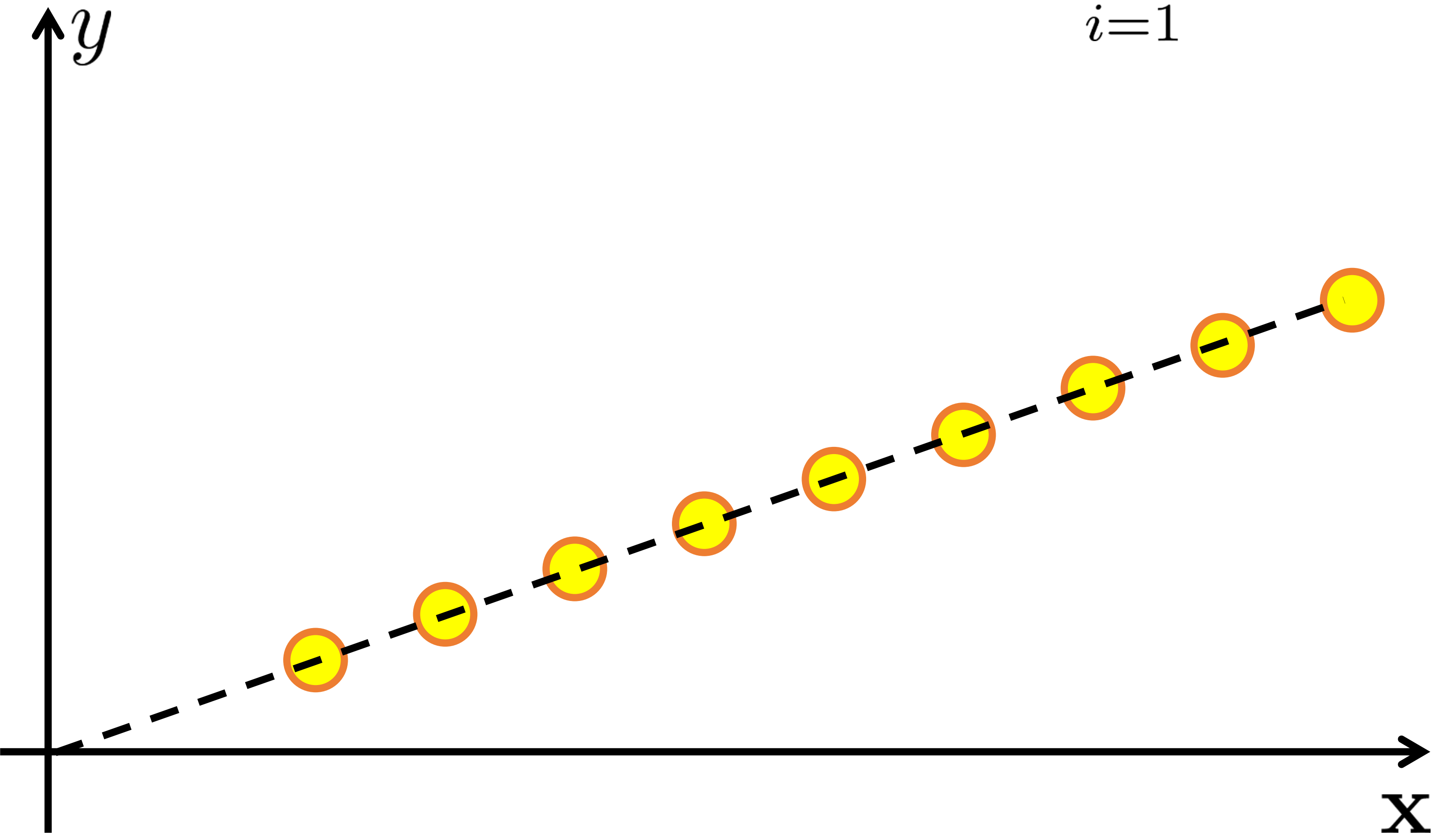
Sparse reg.

Also a sparse reg.

Requires pos. constr.

Entropic reg.

$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$



Regularization

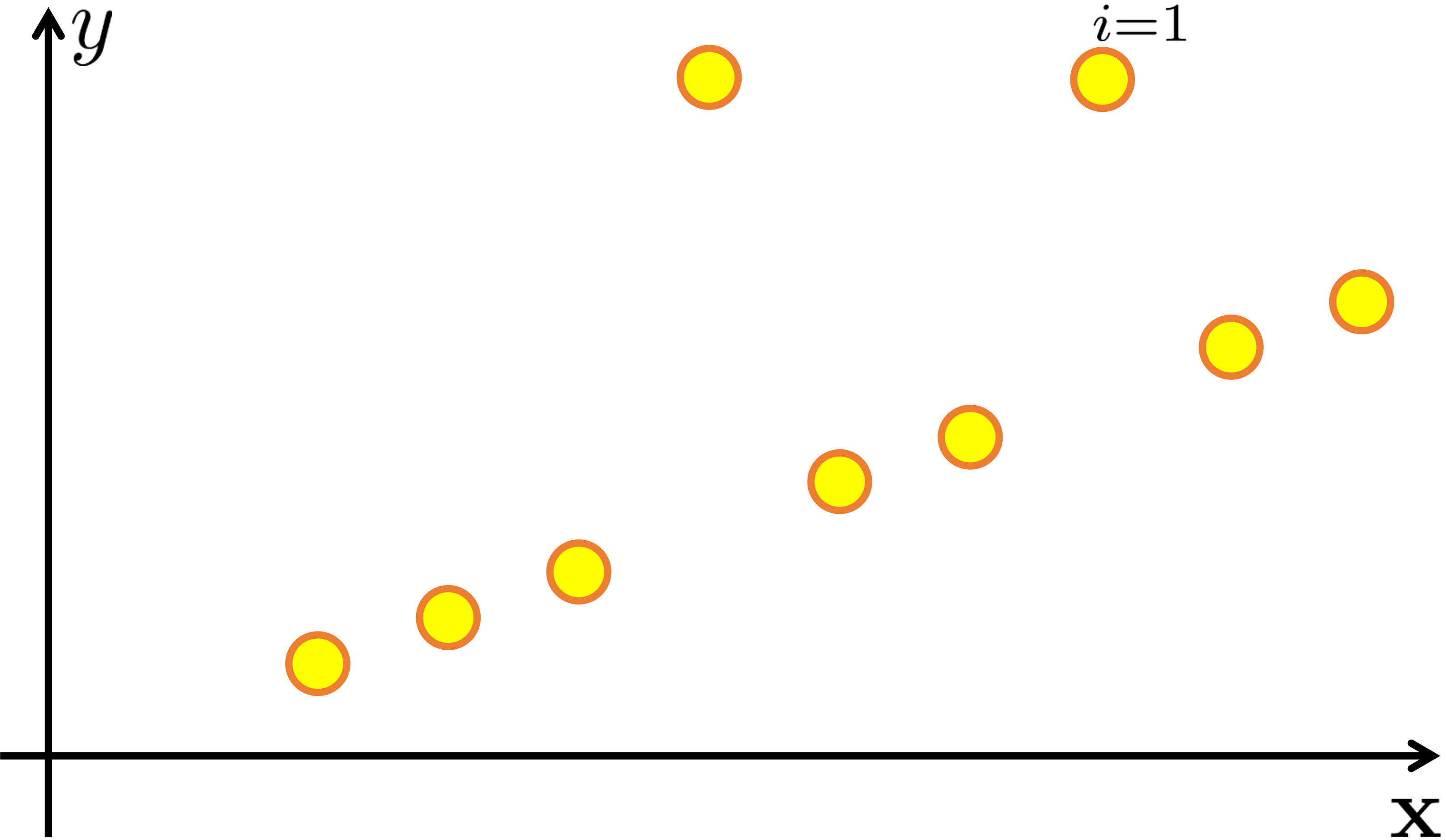
Sparse reg.

Also a sparse reg.

Requires pos. constr.

Entropic reg.

$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$



Regularization

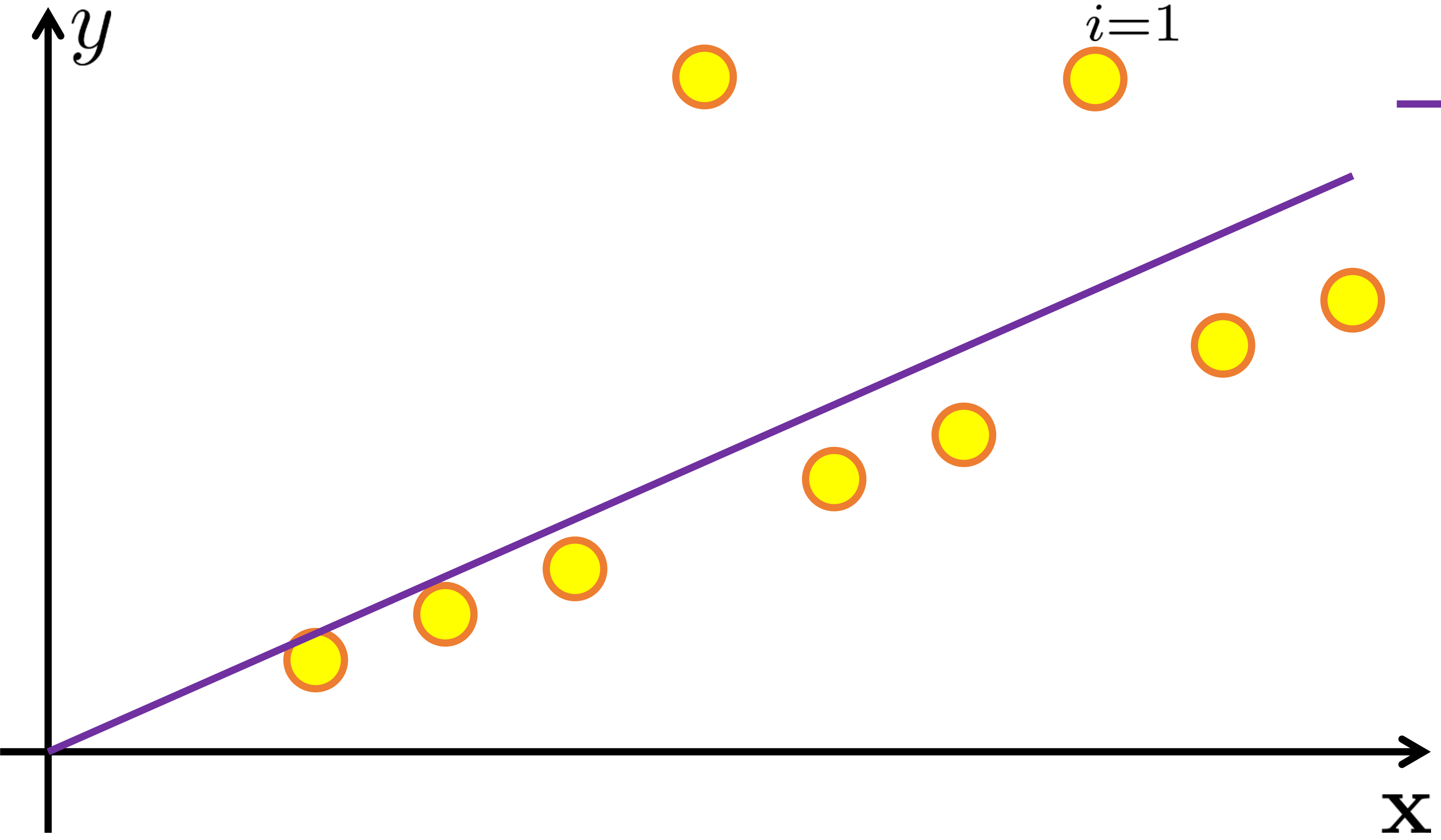
Sparse reg.

Also a sparse reg.

Requires pos. constr.

$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$

Entropic reg.



$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$$

No regularization!

$$\sum\left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

Regularization

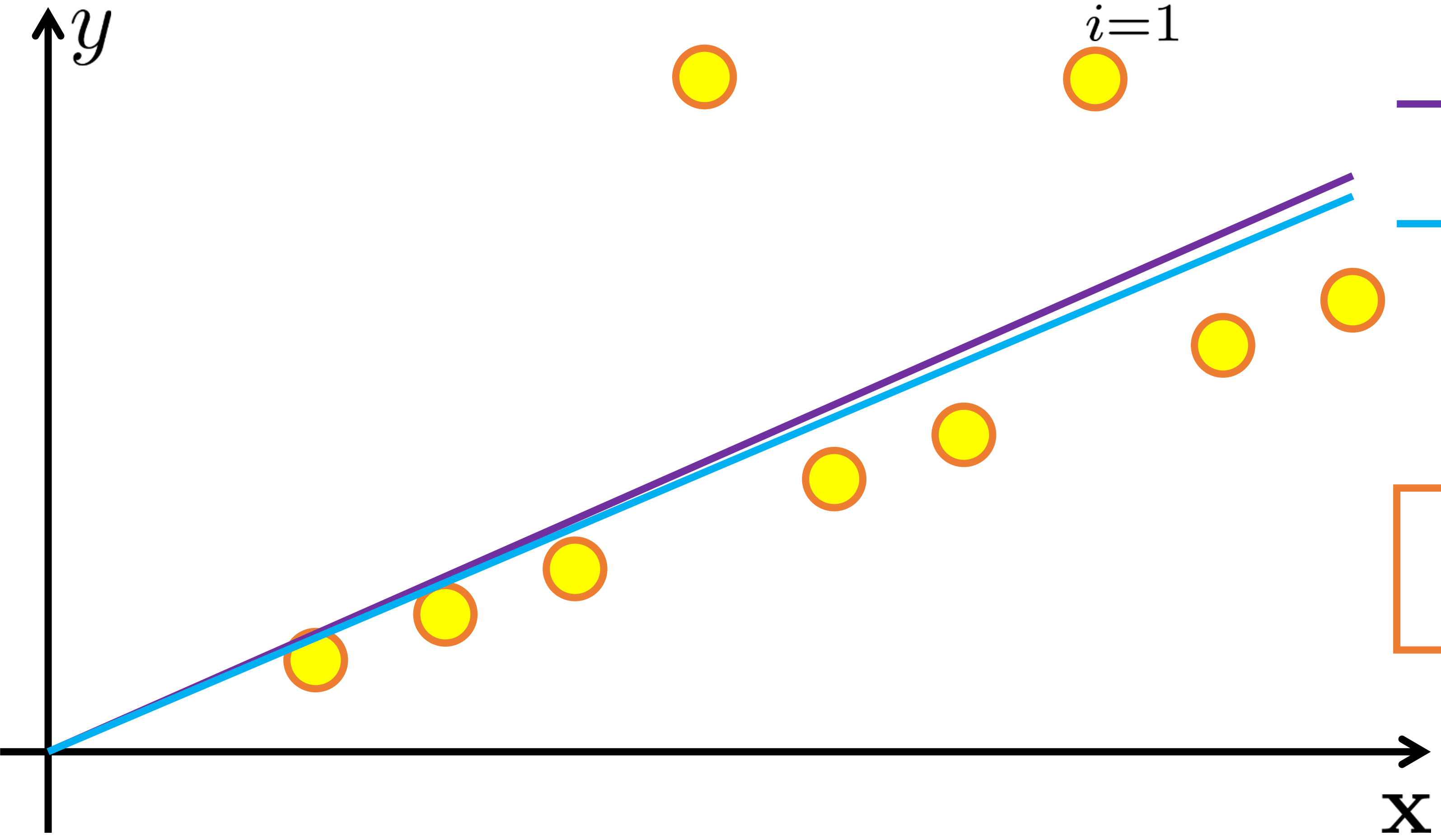
Sparse reg.

Also a sparse reg.

Requires pos. constr.

Entropic reg.

$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$



$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 0.1I)^{-1}\mathbf{X}\mathbf{y}$$

Feeble regularization!

$$\sum\left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + 0.1\|\mathbf{w}\|_2^2$$

Regularization

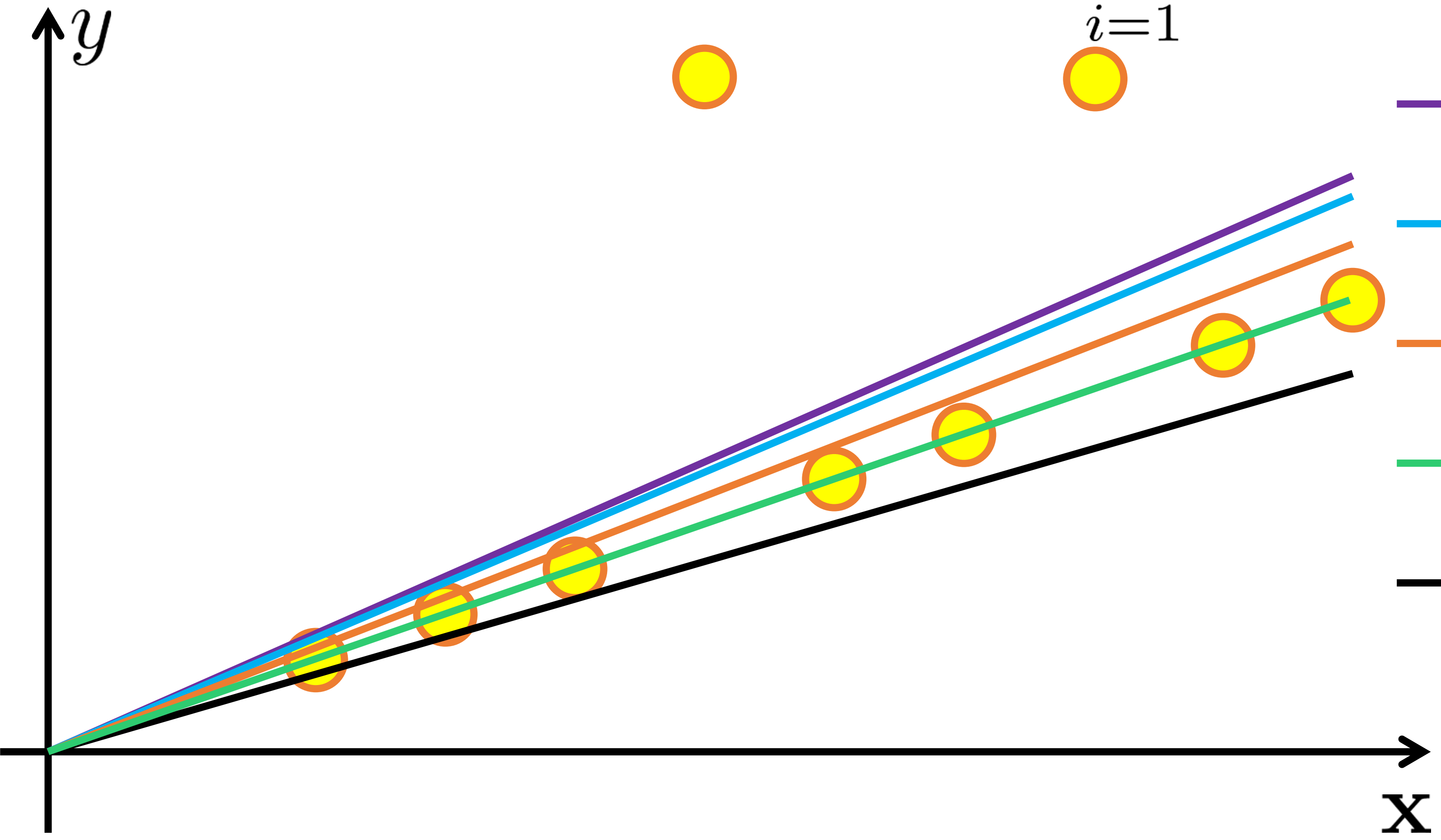
Sparse reg.

Also a sparse reg.

Requires pos. constr.

Entropic reg.

$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$



$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 0.1I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 0.5I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 2I)^{-1}\mathbf{X}\mathbf{y}$$

Regularization

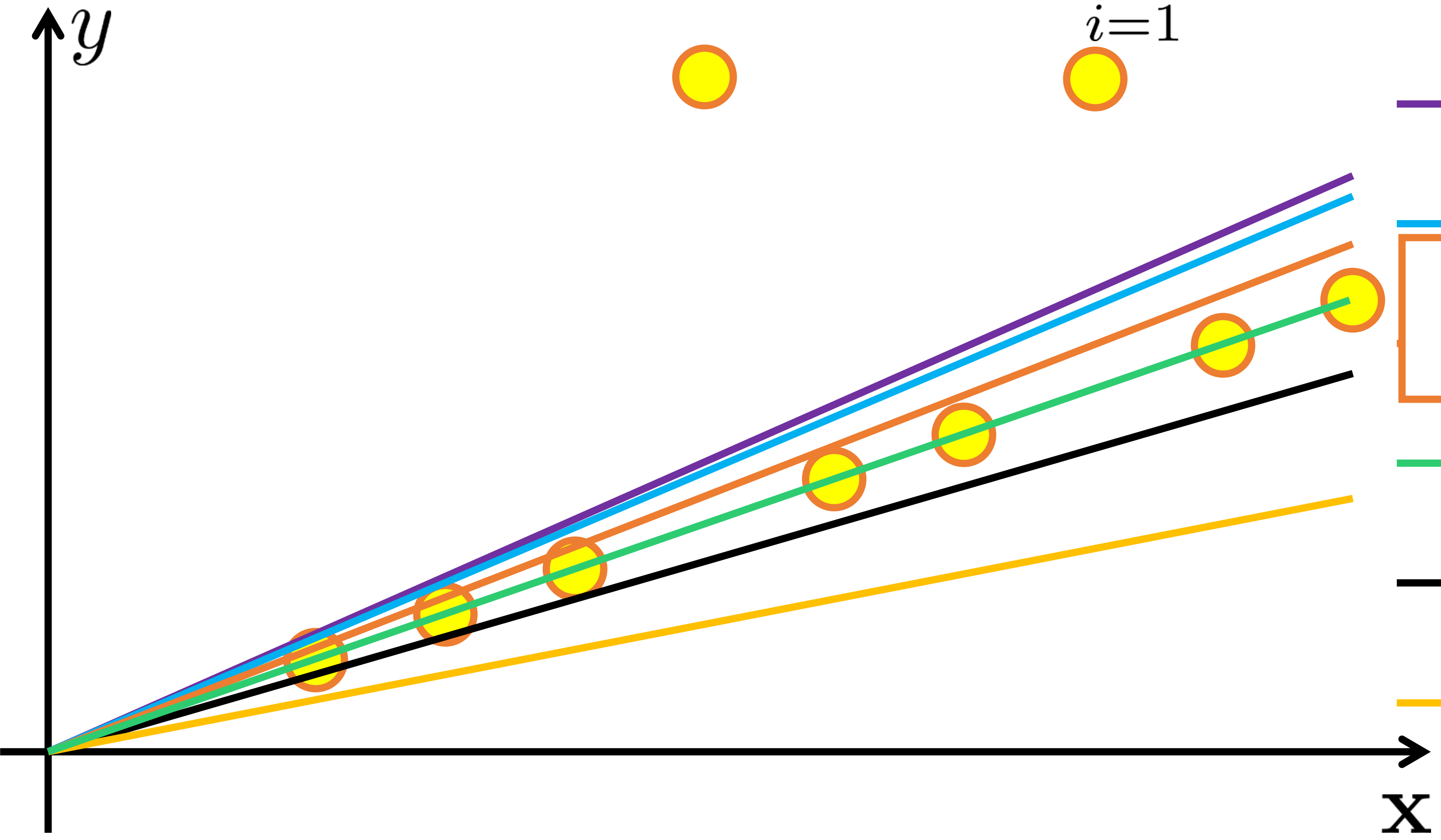
$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$

Sparse reg.

Also a sparse reg.

Requires pos. constr.

Entropic reg.



$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 0.1I)^{-1}\mathbf{X}\mathbf{y}$$

Strong regularization!

$$\sum\left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + 5\|\mathbf{w}\|_2^2$$

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

$$= (\mathbf{X}\mathbf{X}^\top + 2I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 5I)^{-1}\mathbf{X}\mathbf{y}$$

Regularization

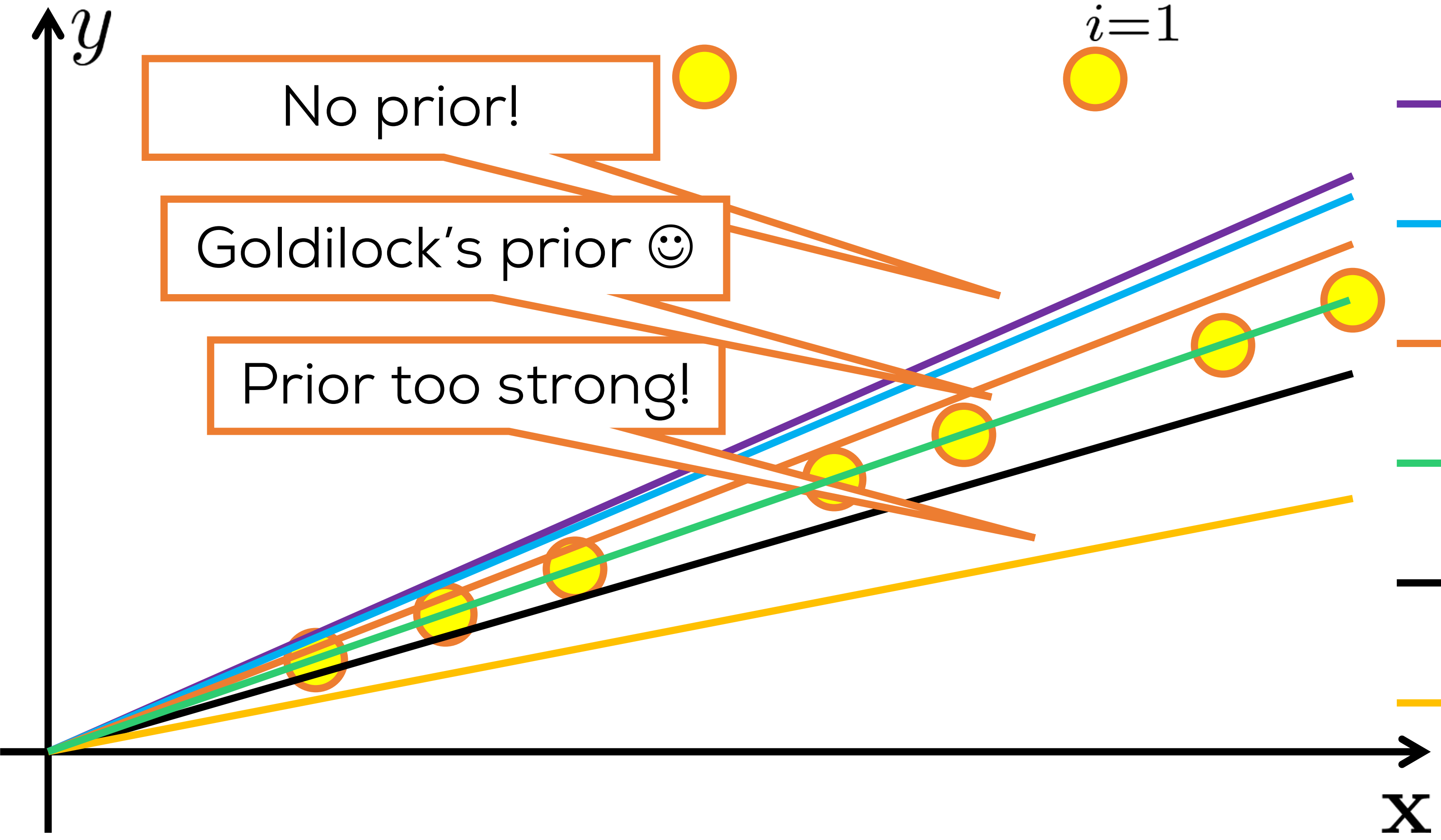
$$\lambda \cdot \|\mathbf{w}\|_2^2 \qquad \lambda \cdot \|\mathbf{w}\|_1 \qquad \lambda \cdot \sum_{i=1}^{d} \mathbf{w}_i \log \mathbf{w}_i$$

Sparse reg.

Also a sparse reg.

Requires pos. constr.

Entropic reg.



$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 0.1I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 0.5I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + I)^{-1}\mathbf{X}\mathbf{y}$$

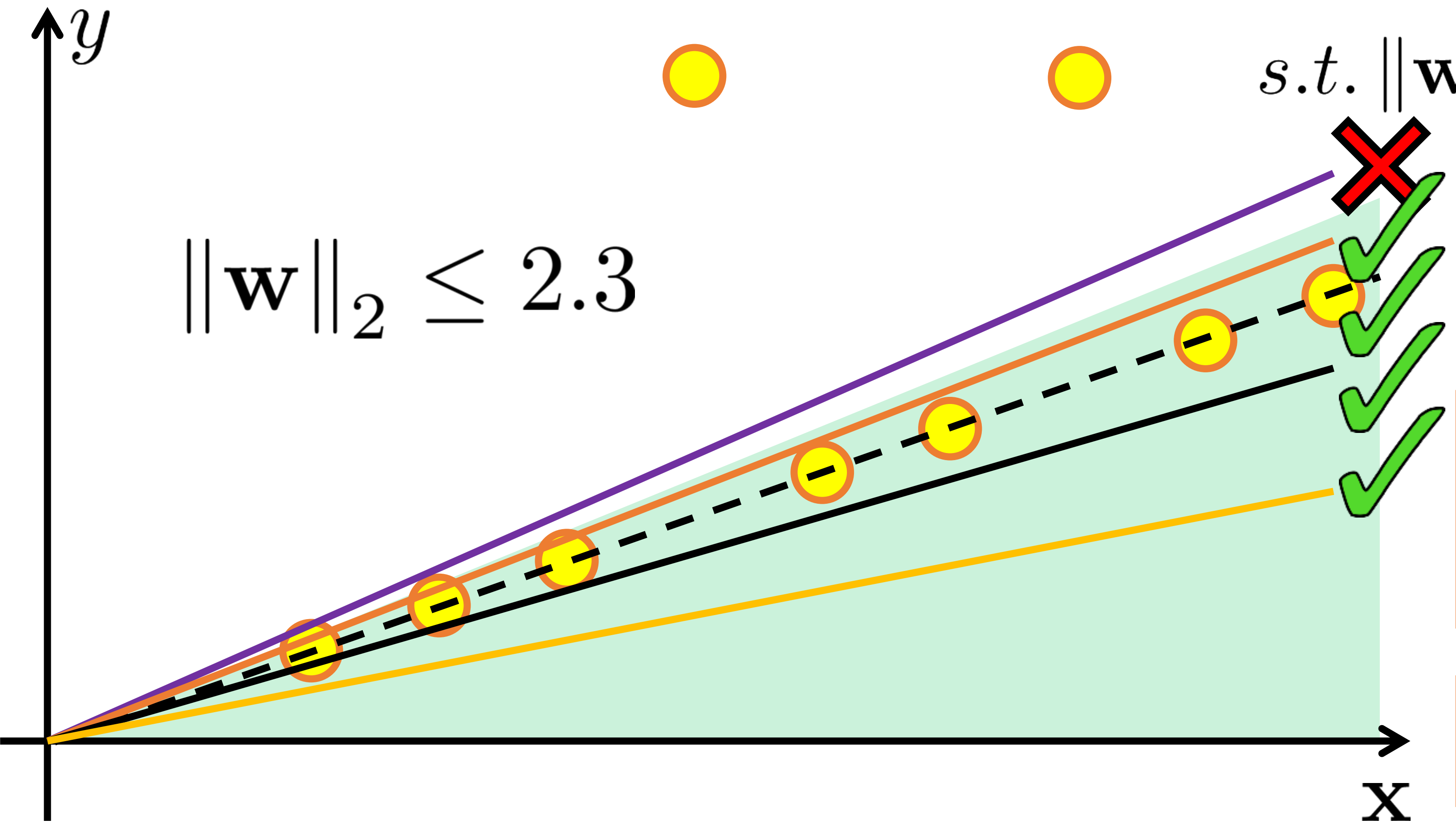
$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 2I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top + 5I)^{-1}\mathbf{X}\mathbf{y}$$

Constrained Optimization

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

$$s.t. \ \|\mathbf{w}\|_2 \leq r$$



Constraint

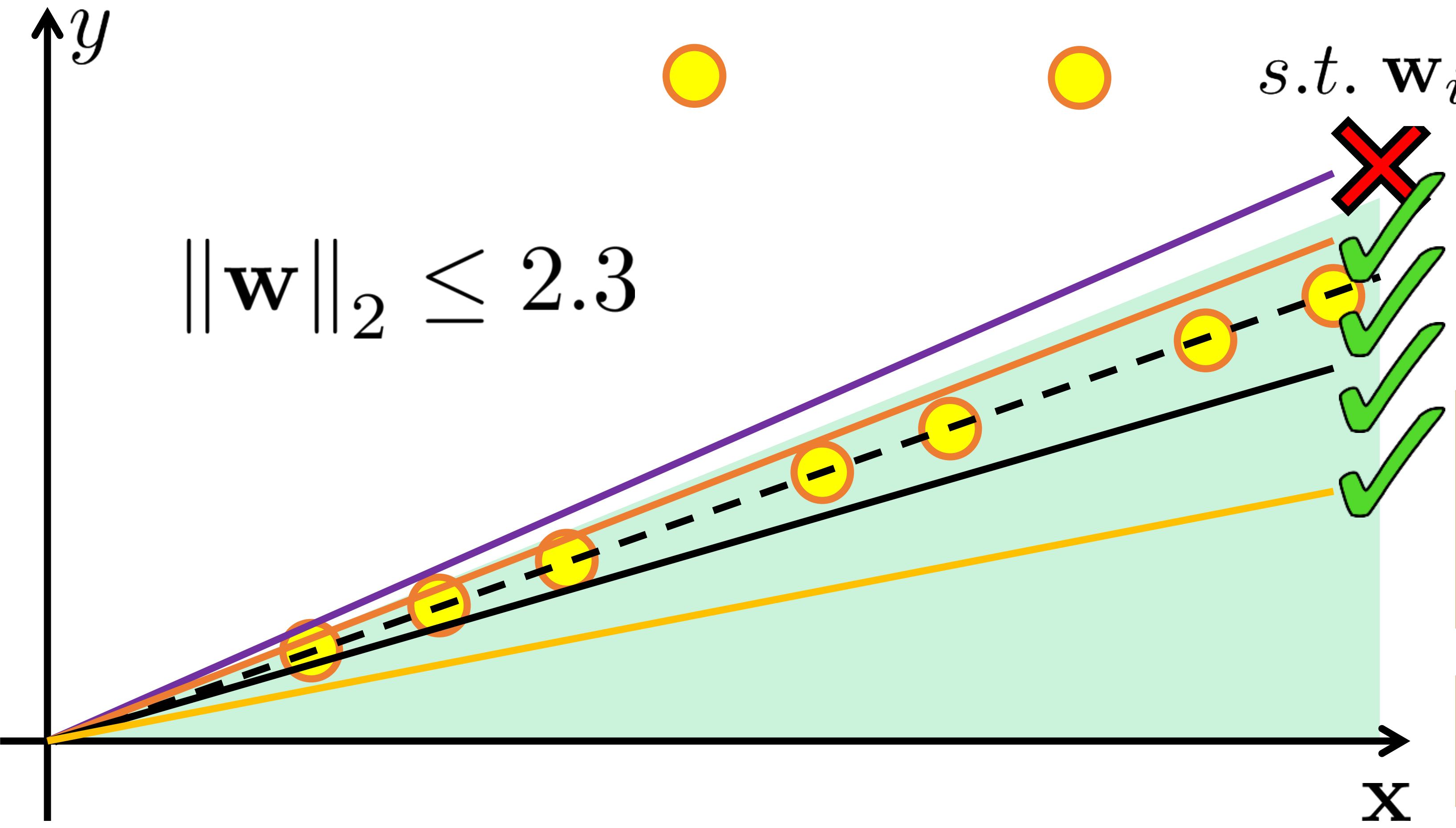
Sparsity constraint
$\|\mathbf{w}\|_1 \leq r, \|\mathbf{w}\|_0 \leq s$

Direct control!

Constrained Optimization

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

$$s.t.\ \mathbf{w}_i \geq 0, \sum \mathbf{w}_i \log \mathbf{w}_i \leq r$$



Constraint

Sparsity constraint
$\|\mathbf{w}\|_1 \leq r, \|\mathbf{w}\|_0 \leq s$

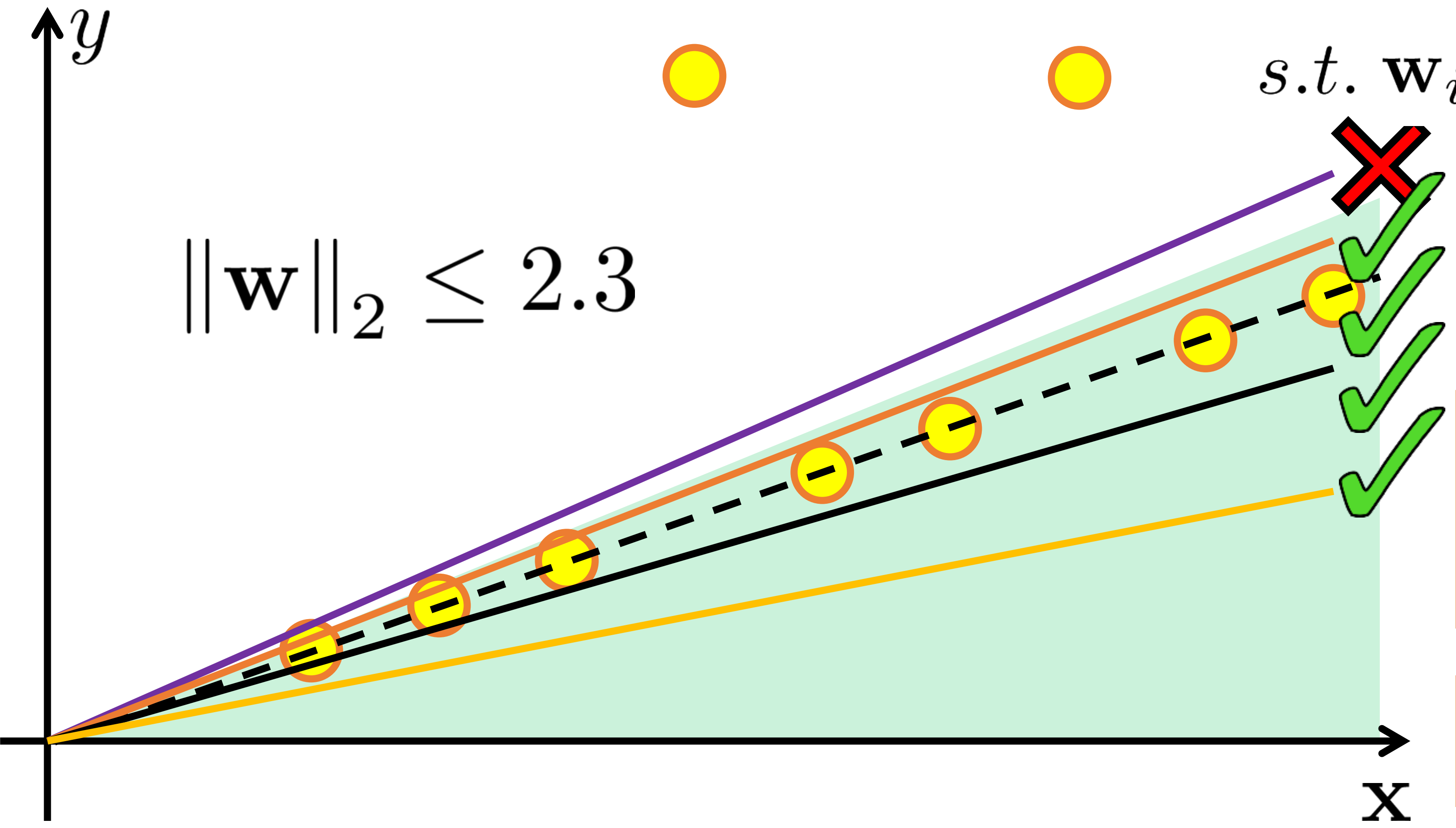
Direct control!

Constrained Optimization

Same power as regularization

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

$$s.t.\ \mathbf{w}_i \geq 0, \sum \mathbf{w}_i \log \mathbf{w}_i \leq r$$



Prior??

Constraint

Sparsity constraint
$\|\mathbf{w}\|_1 \leq r, \|\mathbf{w}\|_0 \leq s$

Direct control!

The FA philosophy

Inductive bias

Non-linear? Wait

- Hypothesis: $\mathbf{x}^i \rightarrow y^i$ can be approximated using a linear function
- Hypothesis: the approximation is captured using loss function ℓ
- Hypothesis: the linear function is "simple"

Small norm, etc.

The FA philosophy

Inductive bias

Non-linear? Wait

- Hypothesis: $\mathbf{x}^i \to y^i$ can be approximated using a linear function
- Hypothesis: the approximation is captured using loss function ℓ
- Hypothesis: the linear function is “simple”

Small norm, etc.

Regression: $\exists \mathbf{w}$, such that $y^i \approx \langle \mathbf{w}, \mathbf{x}^i \rangle$

Binary Classification: $\exists \mathbf{w}$, s.t. $y^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$

Multi Classification: $\exists \{\mathbf{w}^j\}$, s.t. $y^i = \arg\max \langle \mathbf{w}, \mathbf{x}^i \rangle$

Multi-label Classification: $\exists \{\mathbf{w}^j\}$, s.t. $\mathbf{y}^i_j = \text{sign}(\langle \mathbf{w}^j, \mathbf{x}^i \rangle)$

Basket model: $\exists\ \boldsymbol{W} = \{\mathbf{w}^k\}, \boldsymbol{V} = \{\mathbf{v}^k\}$ s.t. $\mathbf{y} = \text{sign}(\boldsymbol{V}\boldsymbol{\alpha}), \boldsymbol{\alpha} = \boldsymbol{W}\mathbf{x}^i$

The FA philosophy

Inductive bias

Non-linear? Wait

- Hypothesis: $\mathbf{x}^i \to y^i$ can be approximated using a linear function
- Hypothesis: the approximation is captured using loss function ℓ
- Hypothesis: the linear function is "simple"

Small norm, etc.

The FA Approach	The PML Approach
• Approximation mechanism	• Generative mechanism
• Linear function	• Linearly parameterized dist.
• Loss function on pred. y	• Likelihood dist. on pred. y
• Reg./constraint on model $\mathbf{w}$	• Prior distribution on model $\mathbf{w}$
• Bayesian FA: Bandit Optimization	• FA-style PML: MAP, MLE

The FA philosophy

- Choose the model that looks "good" on training data
- Akin to choosing the model with high likelihood or posterior
- Empirical Risk Minimization

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

- Empirical Regularized Risk Minimization

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

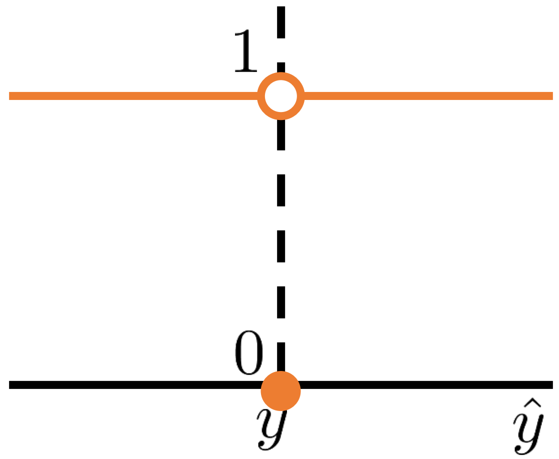
Fantastic FA Formulations

and how to construct them

Regression Loss Functions

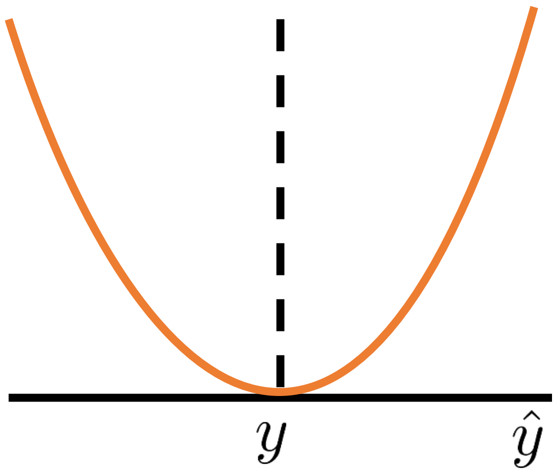
$$\ell_{0-1}(y, \hat{y}) = \mathbb{I}\{y \neq \hat{y}\}$$

0-1 loss



$$\ell_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2$$

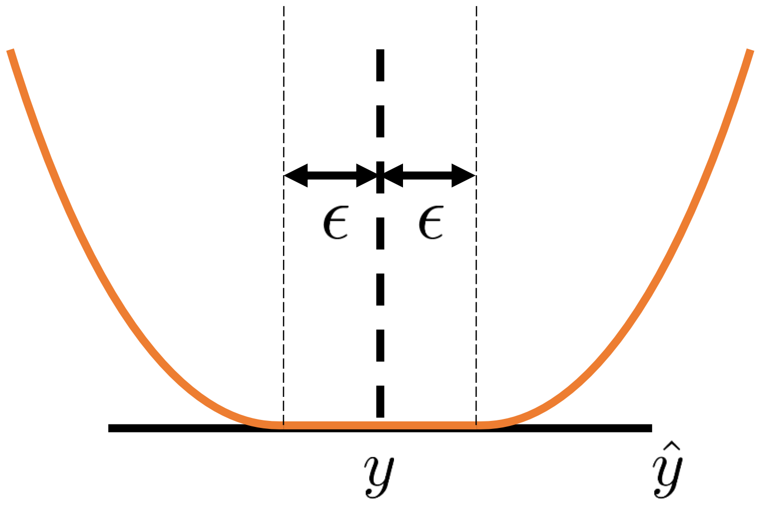
Square loss



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

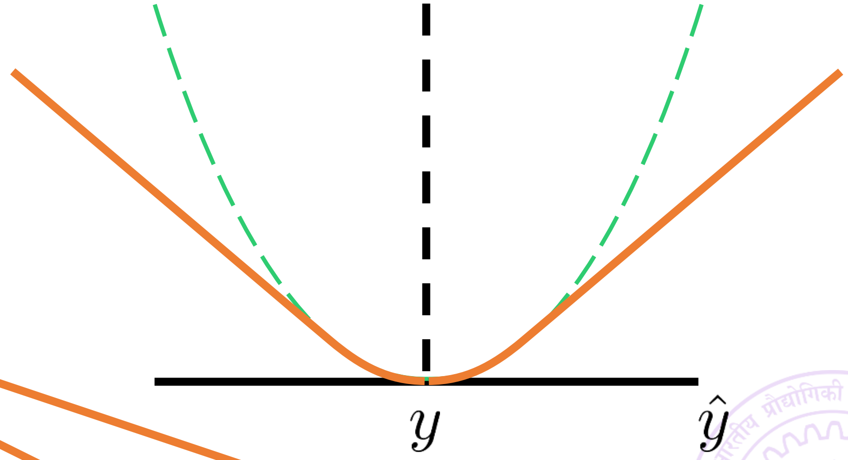
Regression Loss Functions

$$\ell_\epsilon(y, \hat{y}) = \begin{cases} (y - \hat{y} - \epsilon)^2 & \text{if } \hat{y} < y - \epsilon \\ 0 & \text{if } \hat{y} - y \in [-\epsilon, \epsilon] \\ (y - \hat{y} + \epsilon)^2 & \text{if } \hat{y} > y + \epsilon \end{cases}$$



Vapnik's ϵ-insensitive loss

$$\ell_\delta(y, \hat{y}) = \begin{cases} (\hat{y} - y)^2 & \text{if } |\hat{y} - y| \leq \delta \\ \delta \cdot |y - \hat{y}| & \text{if } |\hat{y} - y| \geq \delta \end{cases}$$



Huber loss

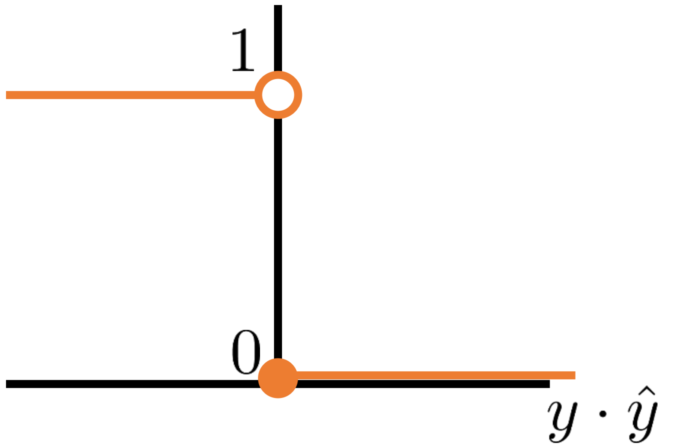
$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Other loss functions for GLM, quantile/ordinal regression

Binary Classification Loss Functions

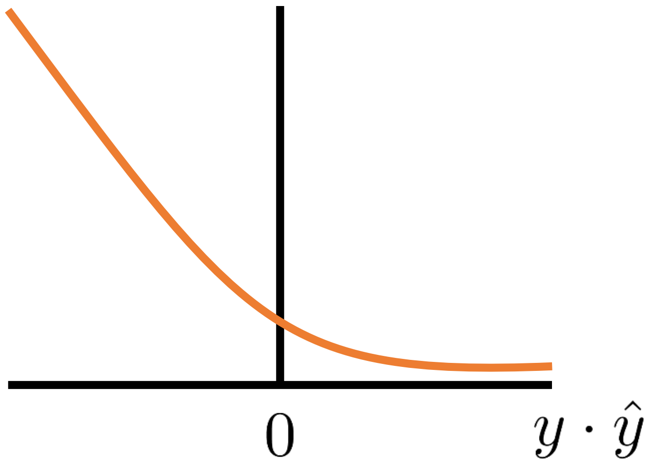
$$\ell_{0-1}(y, \hat{y}) = \mathbb{I}\{y \neq \text{sign}(\hat{y})\}$$

0-1 loss



$$\ell_{\log}(y, \hat{y}) = \log(1 + \exp(-y \cdot \hat{y}))$$

Logistic loss

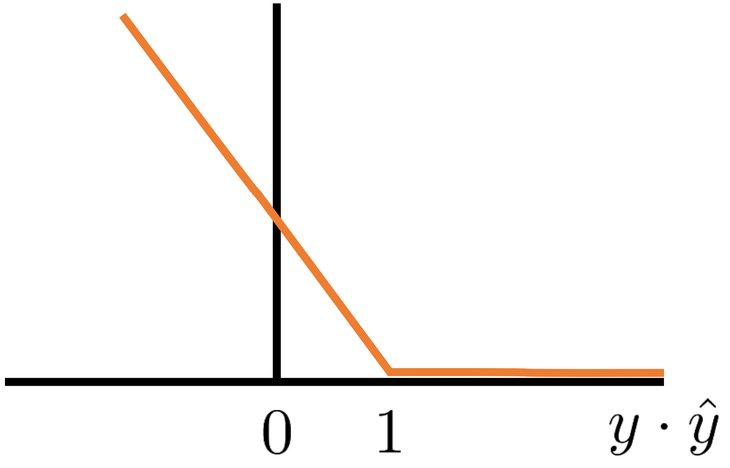


$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Binary Classification Loss Functions

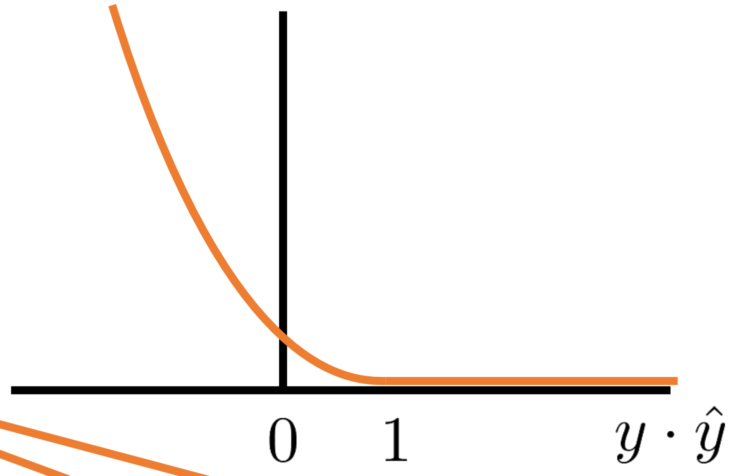
$$\ell_{\text{hinge}}(y, \hat{y}) = [1 - y \cdot \hat{y}]_+$$

Hinge loss



$$\ell_{\text{sq-hinge}}(y, \hat{y}) = [1 - y \cdot \hat{y}]_+^2$$

Squared Hinge loss



Other loss functions for imbalanced problems

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

Looking into the Hinge Loss

Can use non-linear functions too!

Binary Classification

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

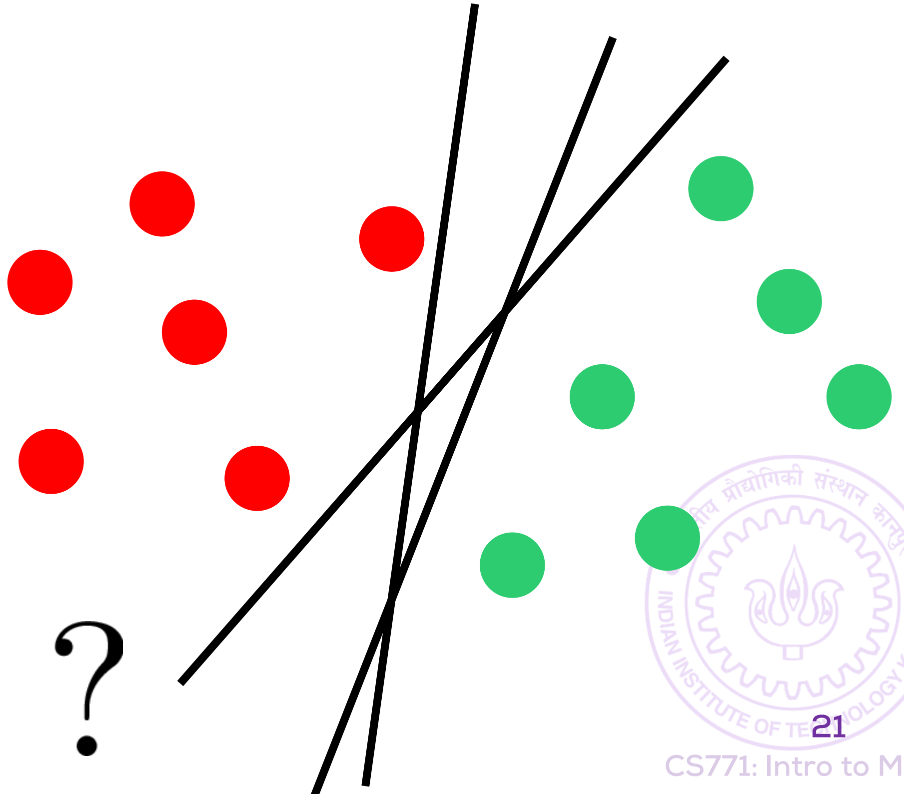
Looking into the Hinge Loss

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!



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!

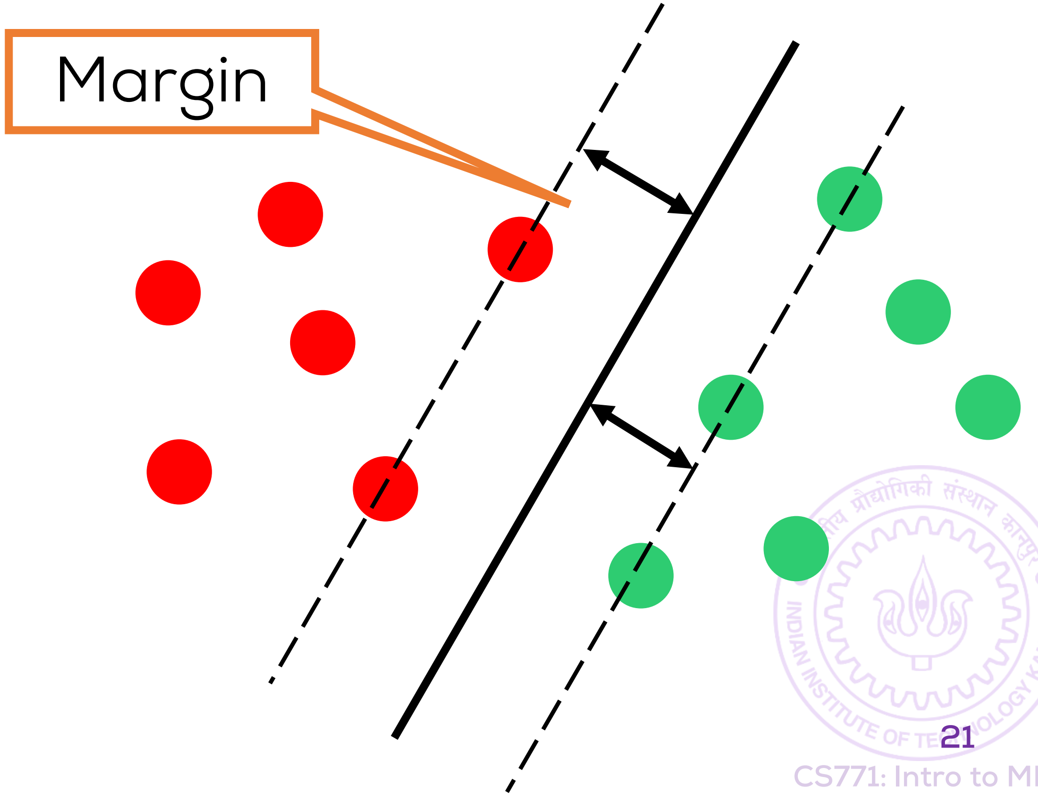
$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

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

Regularization parameter

Margin

Why 1?



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

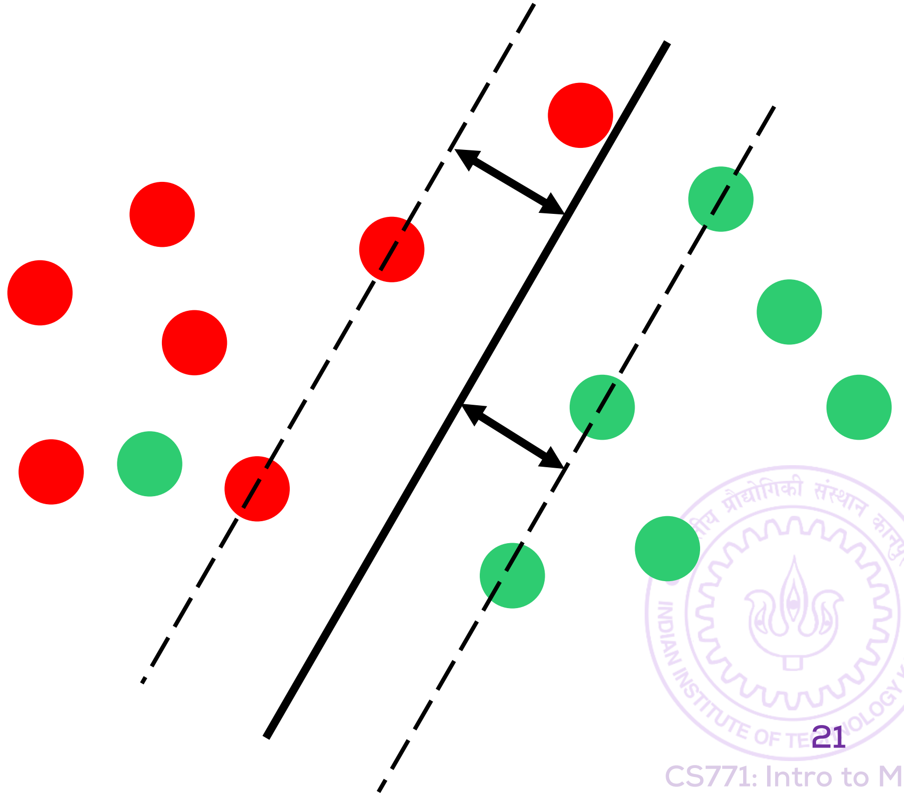
Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

Slack variable

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

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

$$\xi_i \geq 0$$



Looking into the Hinge Loss

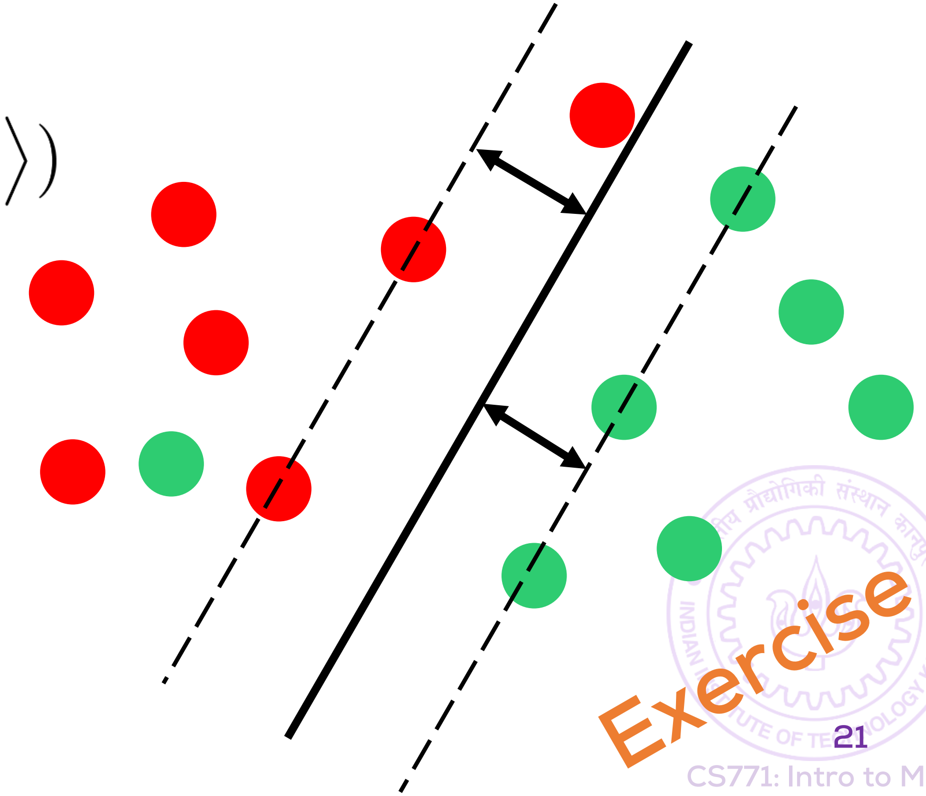
Binary Classification

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

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

Looking into the Hinge Loss

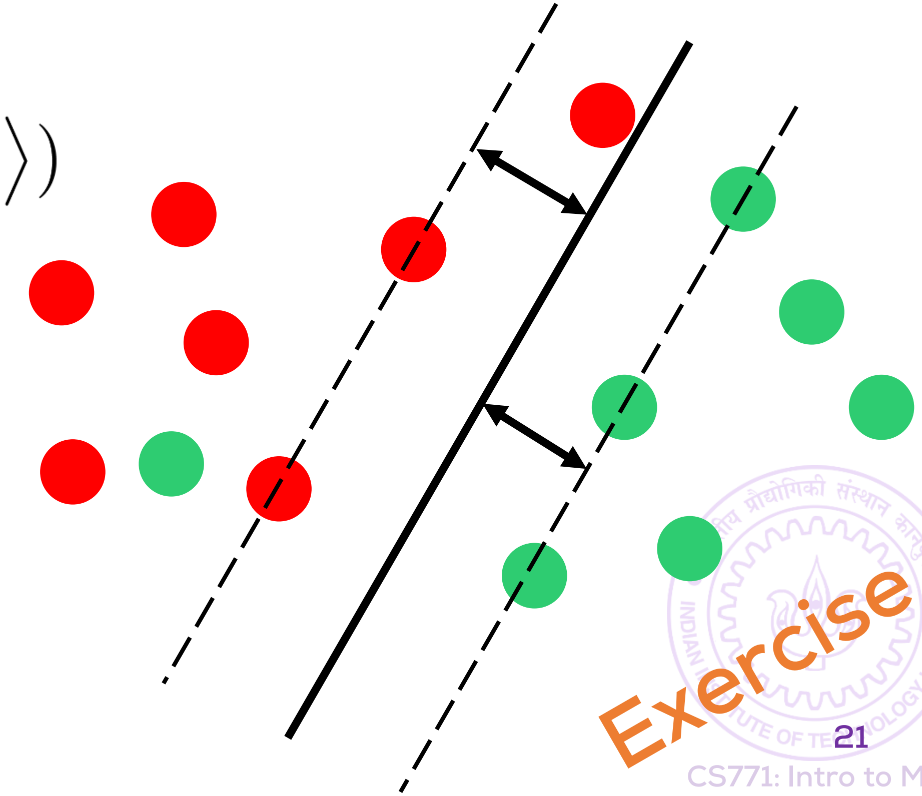
Binary Classification

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

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

Looking into the Hinge Loss

Binary Classification

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

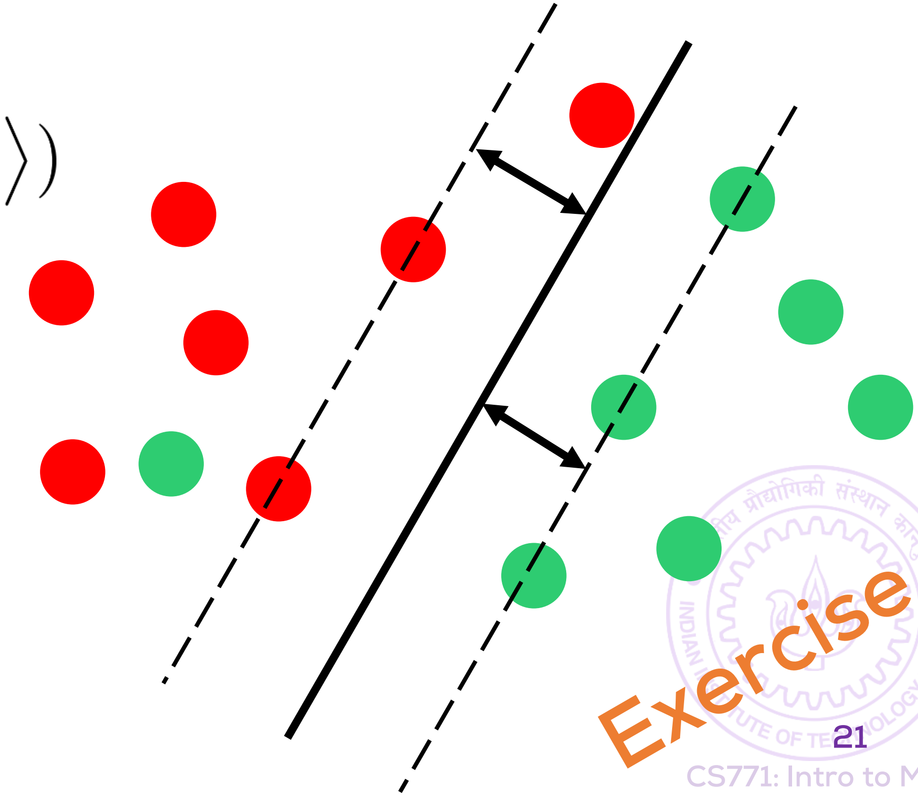
Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

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

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

Large Margin Classifier

SVM



Exercise

Please give your Feedback

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