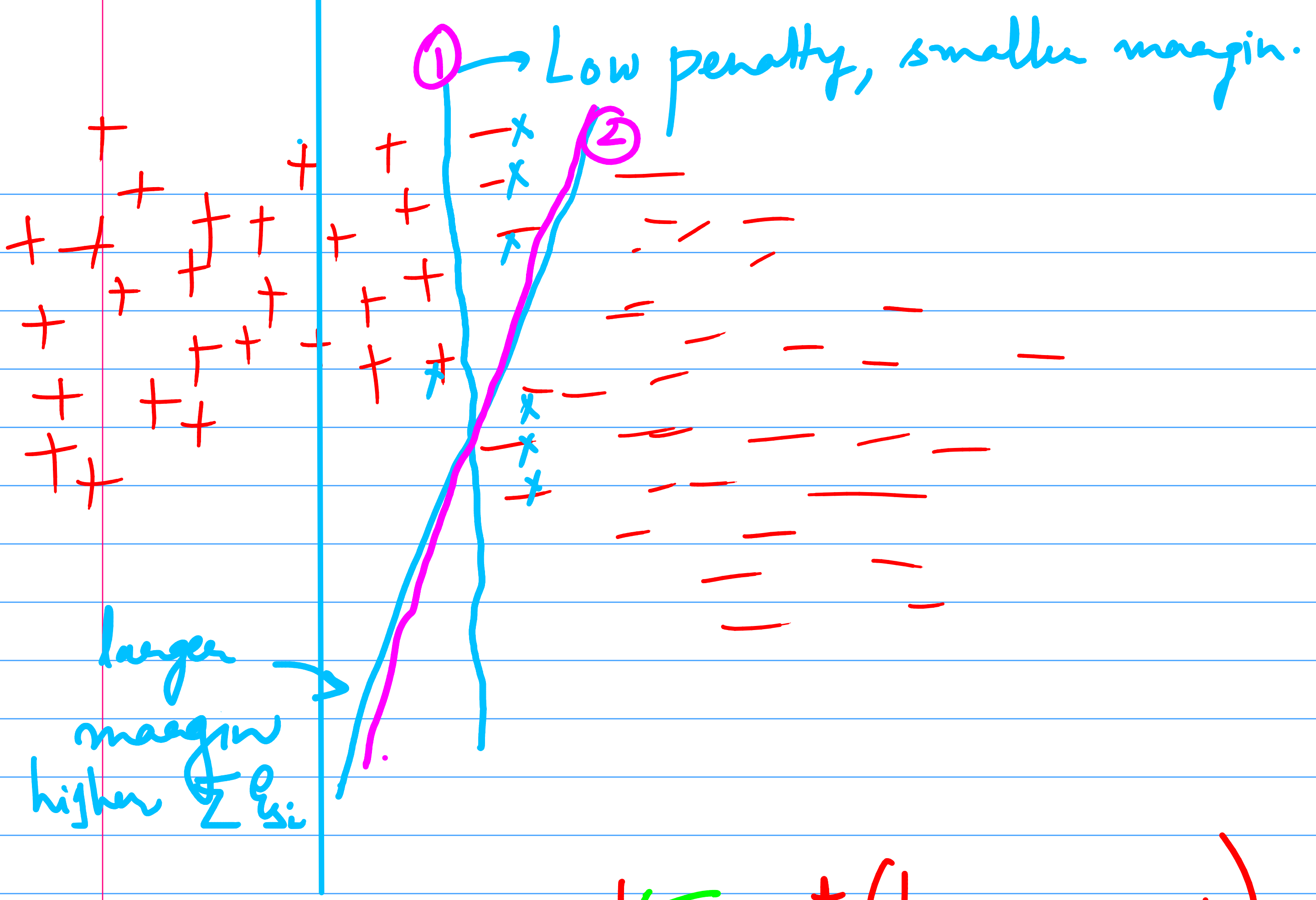
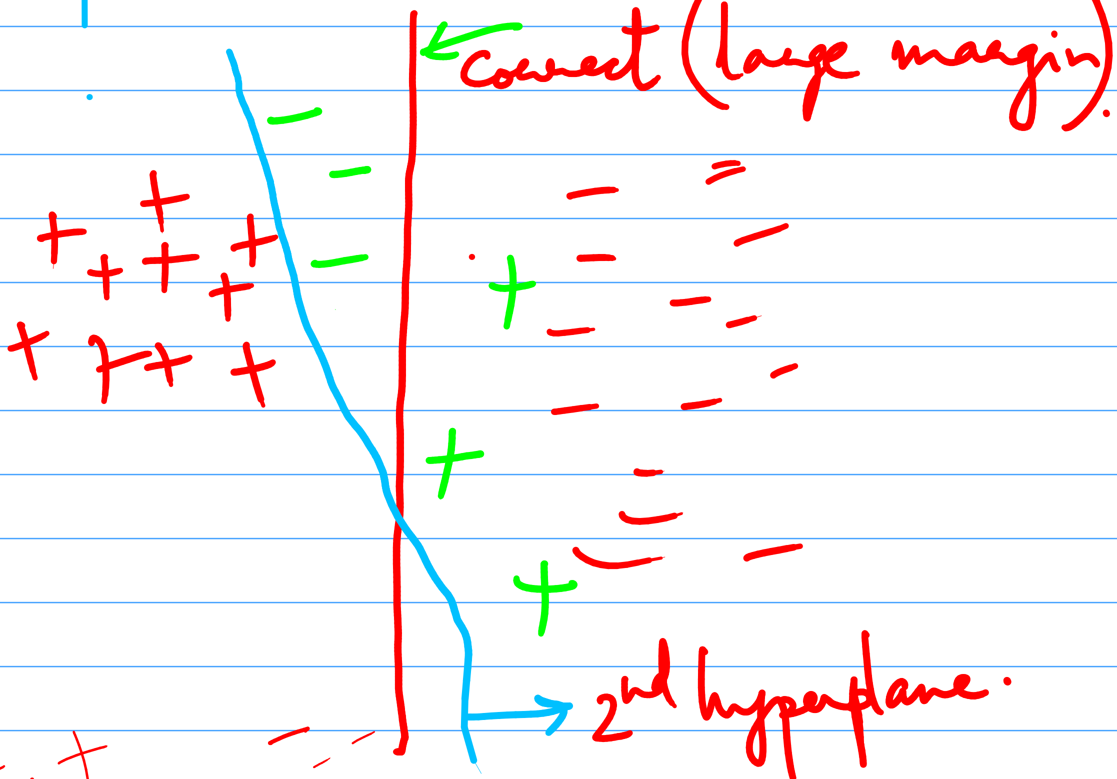
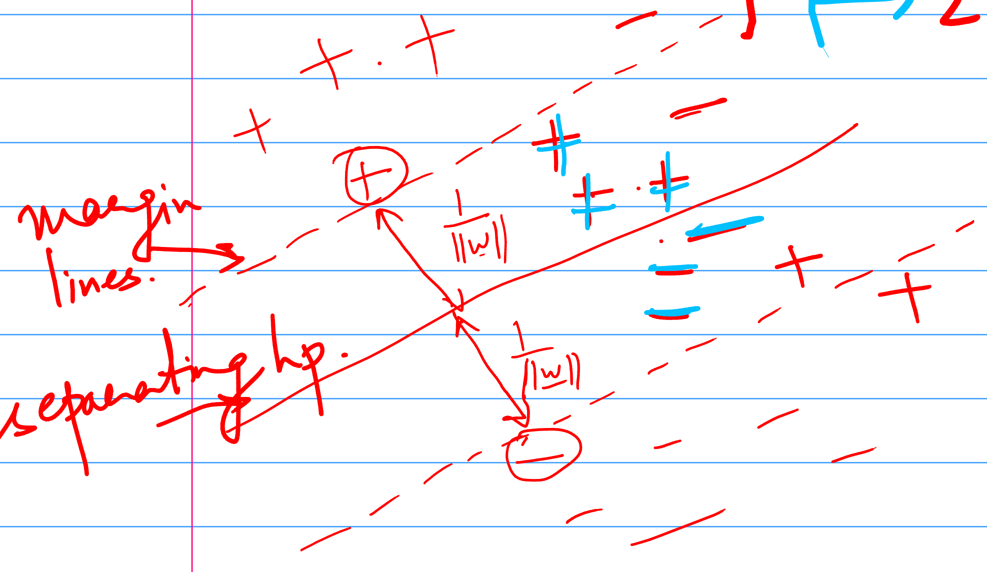

1
Low penalty, smaller margin.
2
larger
margin
higher $\sum \xi_i$


correct (large margin).
2nd hyperplane.

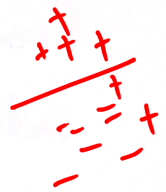

margin
lines.
separating hp.
$\frac{1}{||w||}$
$\frac{1}{||w||}$


Hard SVM
small
margin

Soft SVM Rule

Soft SVM is a relaxation of the Hard SVM rule.
It can be applied even if the training set is not linearly separable.

In case of separable data, we expect all examples to satisfy the constraints $\underbrace{y_i(\langle \underline{w}, x_i\rangle + b)}_{LHS} \geq 1$. $\frac{1}{\|\underline{w}\|}$



However, when the data is not linearly separable, the LHS of the constraint may fall short of 1 by a quantity $\xi_{s_i} > 0$ $\quad \xi_{s_i} \ \forall i \quad$ m such unknowns

If we revise the constraints as

$$\underbrace{y_i(\langle \underline{w}, x_i\rangle + b)}_{LHS} \geq 1 - \xi_{s_i} \quad \text{RHS.}$$

then every example can be assigned a ξ_{s_i} value such that for $\xi_{s_i} \geq 0$ all the m constraints are satisfied.

The quantity ξ_{s_i} is a slack variable.

The soft SVM learning rule is formulated as

$$\min_{\underline{w},\, b,\, \underline{\xi_s}} \left(\underset{\text{fixed}}{\lambda} \|\underline{w}\|^2 + \underbrace{\frac{1}{m} \sum_{i=1}^{m} \xi_{s_i}}_{\text{penalty term}} \right)$$

such that $\forall i \quad y_i(\langle \underline{w}, x_i\rangle + b) \geq 1 - \xi_{s_i}$

Primal formulation

not adjusted / learned during training

λ is a hyper parameter which controls the tradeoff between the two terms. but they are assigned a value before training

Output of this optimization is $\underline{w}$, b and ξ_i

Fix $\underline{w}, b$ and consider minimization over ξ_s

$\xi_{s_i} \geq 0$

Every ξ_{s_i} is non-negative, so we can minimize them individually.

$\frac{1}{m}\sum \xi_{s_i}$ should be min

Consider $y_i(\langle \underline{w}, \underline{x_i}\rangle + b) \geq 1 - \xi_{s_i}$

If LHS is already greater than 1, then

$\min \xi_{s_i} = 0$

But if LHS is less than 1

then to ensure that LHS $\geq$ RHS,

RHS should be decreased by increasing ξ_{s_i}

$$\xi_{s_i} \geq 1 - y_i\underbrace{(\langle \underline{w}, \underline{x_i}\rangle + b)}_{\hat{y}_i}$$

$y_i\hat{y}_i \geq 1 - \xi_{s_i}$

$\xi_{s_i} \geq 1 - y_i\hat{y}_i$

$$\min \xi_{s_i} = 1 - y_i(\langle \underline{w}, \underline{x_i}\rangle + b)$$

$\xi_{s_i} = 1 - y_i\hat{y}_i$

Therefore, as a result of minimization,

either $\xi_{s_i} = 0$ or $\xi_{s_i} = 1 - y_i(\langle \underline{w}, \underline{x_i}\rangle + b)$ $\qquad \xi_i \geq 0$

$$\Rightarrow \xi_{s_i} = \max(0, 1 - y_i(\langle \underline{w}, \underline{x_i}\rangle + b))$$

for no penalty $\xi_{s_i} = \max(0, -ve)$

there is penalty $\xi_{s_i} = \max(0, +ve)$

??

This formulation is similar to that of Hinge Loss.

Soft SVM rule
$$\min_{\underline{w}, b, \underline{\xi_s}} \left(\underbrace{\lambda \|\underline{w}\|^2}_{\text{Regularization term}} + \underbrace{\frac{1}{m} \sum_{i=1}^{m} \xi_{s_i}}_{\text{Hinge loss}} \right)$$

^ raised to the power

_ subscript

min(w,b,ξs) λ||w||^2

^2.

1/m sum_{i=1}^{m}. ξ_si

Overall, the soft SVM learning rule can be considered as regularized loss minimization.

Even the linear regression machine can be regularized by including a regularization term to the loss function.

$$\underbrace{SSD}_{\text{sum of squared difference.}} + \underbrace{\lambda \|\underline{w}\|^2}_{\text{Regularization term}}$$

This is called as Tikhonov regularization.

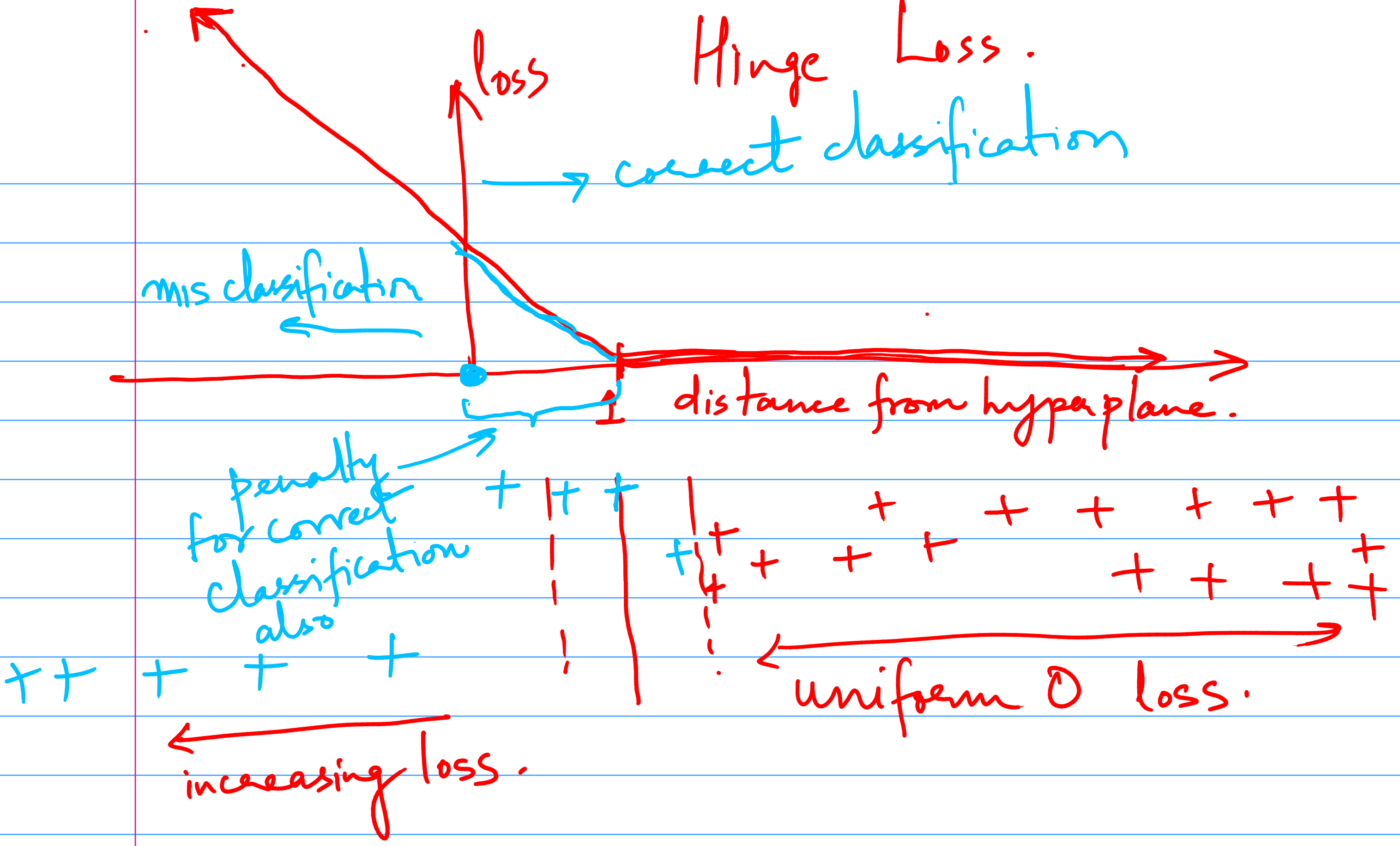
Implementing soft SVM using Stochastic Gradient Descent

$$\min_{\underline{w}} \left(\frac{\lambda}{2}\|\underline{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i \langle \underline{w}, \underline{x_i} \rangle\}}_{L_s(\underline{w})} \right) = L_s(\underline{w})$$

$$\min_{\underline{w}} f(\underline{w})$$

where $f(\underline{w}) = \frac{\lambda}{2} \|\underline{w}\|^2 + L_s(\underline{w})$

Hinge Loss.



Linear Regression (squared error loss)