??

This formulation is similar to that of Hinge Loss.

λ: C parameter

Soft SVM rule

$$\min_{\underline{w}, b, \underline{\xi}} \left(\lambda \|\underline{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_{s_i} \right).$$

λ → tradeoff coefficient

$\lambda \|\underline{w}\|^2$: Regularization term

$\frac{1}{m} \sum_{i=1}^{m} \xi_{s_i}$: Hinge loss

min(w,b,ξ) λ||w||^2

1/m sum_{i=1}^{m} ξi

^ raised to the power

_ subscript

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.

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

Regularization helps in reducing the complexity of a model.

This is called as Tikhonov regularization.

Implementing soft SVM using Stochastic Gradient Descent

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

Hinge loss

sample empirical risk

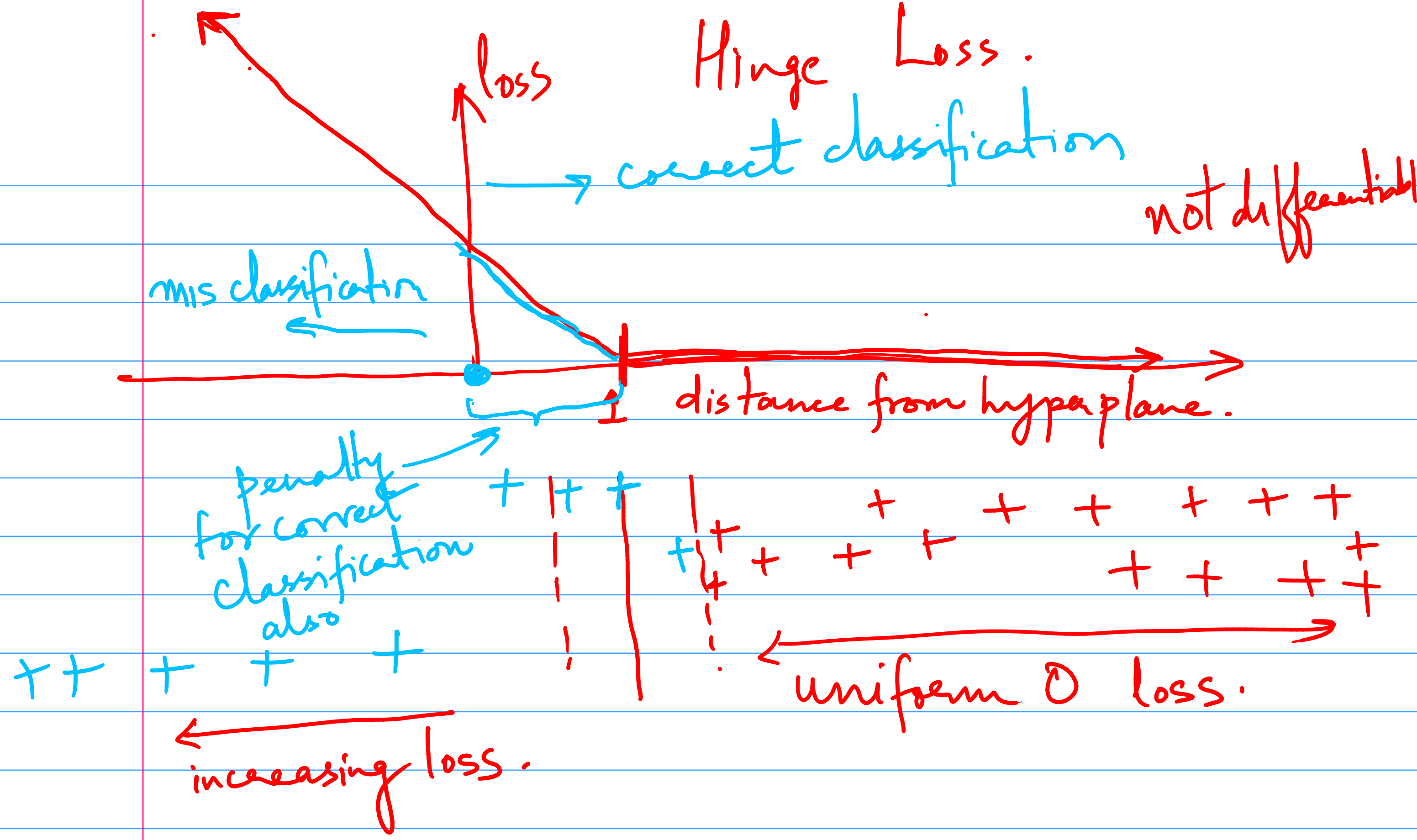
$$\min_{\underline{w}} f(\underline{w})$$

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

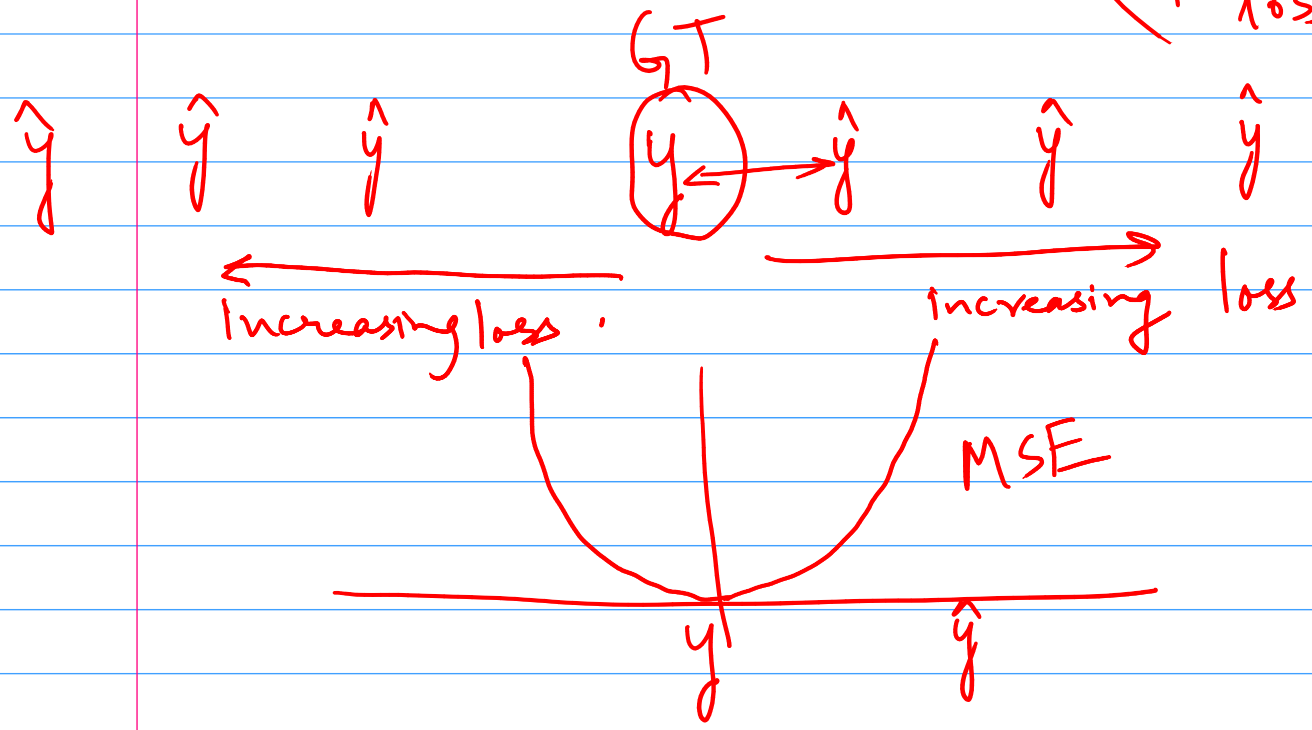
Minimize the true risk ← SGD minimizes.

expected direction is opposite to the gradient vector

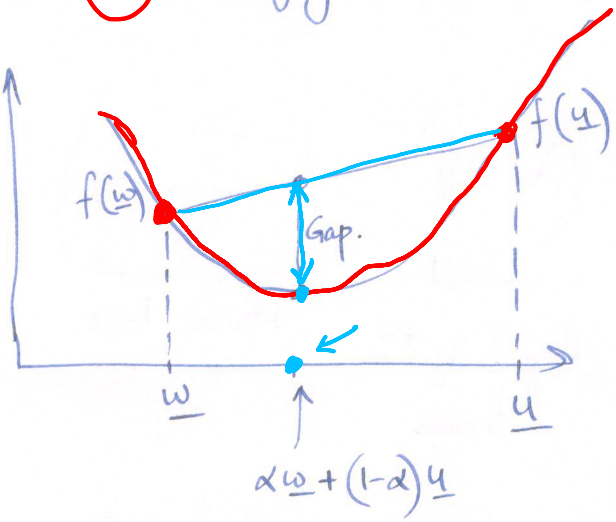
Hinge Loss.



Linear Regression (squared error loss)



f is a λ strongly convex function



For λ strongly convex functions, this $\boxed{\text{gap} \geq \left(\frac{\lambda}{2}\right)\alpha(1-\alpha)\|\underline{u}-\underline{w}\|^2}$

Gradient descent can be performed efficiently for λ strongly convex functions

parameter update $\underline{w}^{(t)} \leftarrow \underline{w}^{(t-1)} - \eta \Delta$ set $\boxed{\eta = \frac{1}{\lambda t}}$

gradient → random SGD..

t : iteration number

Gradient of f at $\underline{w}^{(t)}$

Pick z randomly from training set

Loss for example z $\ell(\underline{w}^{(t)}, z)$

$L_S(\underline{w})$: loss over a sample S

$\ell(\underline{w}^{(t)}, z)$: loss over a single example z.

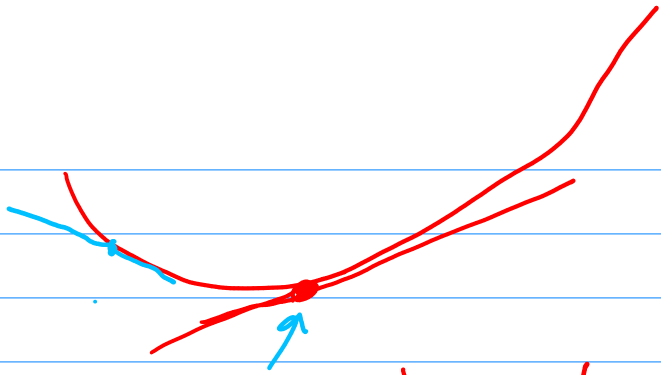
strongly convex function. $\boxed{f(\underline{w}) = \frac{\lambda}{2}\|\underline{w}\|^2 + L_S(\underline{w})}$

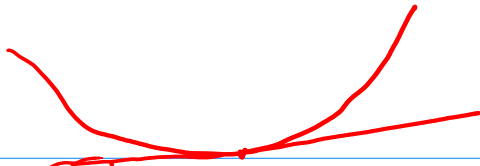
$\ell(\underline{w}^{(t)}, z)$

The loss $\ell(\underline{w}^{(t)}, z)$ may not be differential.

But there can be several subgradient vectors at the point $\underline{w}^{(t)}$ for the function ℓ.

The set of such subgradient vectors is called as the differential set denoted as $\partial \ell(\underline{w}^{(t)}, z)$





tangent vector.

subgradient vectors.

differential set $\partial f(w^{(t)})$.

is the set of all <u>subgradient vectors</u>

Let $\underline{v_t}$ be one such subgradient vector of the loss function $l(\underline{w}^{(t)}, z)$ at point $\underline{w}^{(t)}$

subgradient vector. $\quad \underline{v_t} \in \underbrace{\partial l(\underline{w}^{(t)}, z)}$ differential set of function l

Since $f(\underline{w}) = \underbrace{\frac{\lambda}{2}\|\underline{w}\|^2} + \underbrace{L_S(\underline{w})}$

one of the subgradient vectors of $f(\underline{w})$ at $\underline{w}^{(t)}$ is

subgradient vector $\quad \underbrace{\lambda \underline{w}^{(t)}} + \underline{v_t} \in \underbrace{\partial f(\underline{w}^{(t)})}$

$\therefore$ the SGD (Stochastic Gradient Descent) update step

$$\underline{w}^{(t+1)} = \underline{w}^{(t)} - \eta \underbrace{\left(\lambda \underline{w}^{(t)} + \underline{v_t}\right)}$$

subgradient direction of function f.

For a strongly convex function, $\underline{\eta = \frac{1}{\lambda t}}$

$$\boxed{\therefore \quad \underline{w}^{(t+1)} = \underline{w}^{(t)} - \frac{1}{\lambda t}\left(\lambda \underline{w}^{(t)} + \underline{v_t}\right)}$$

$$= \left(1 - \frac{1}{t}\right)\underline{w}^{(t)} - \frac{1}{\lambda t}\underline{v_t}$$

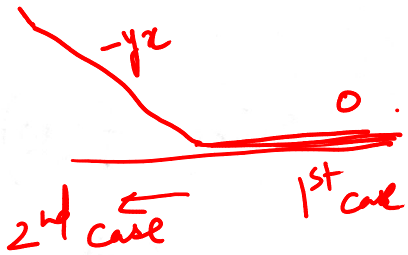
$$= \left(\frac{t-1}{t}\right)\underline{w}^{(t)} - \frac{1}{\lambda t}\underline{v_t}$$

$$= \left(\frac{t-1}{t}\right)\left(\frac{t-2}{t-1}\underline{w}^{(t-1)} - \frac{1}{\lambda(t-1)}\underline{v_{t-1}}\right) - \frac{1}{\lambda t}\underline{v_t}$$

$$\underline{w}^{(t+1)} = -\frac{1}{\lambda t}\left(\sum_{j=1}^{t} \underline{v}_j\right)$$

$\underline{v}_j$ is the subgradient of the loss function at $\underline{w}^{(j)}$

$$= \begin{cases} \underline{0} & \text{1st case} \quad \text{if } y\langle \underline{w}^{(j)}, \underline{x}\rangle \geq 1 \\ -y\underline{x} & \text{otherwise} \quad \text{2nd case} \end{cases}$$



Hinge loss $= \max(0, 1 - y\langle \underline{w}, \underline{x}\rangle)$.

SGD for solving soft SVM

$$\boxed{sum^{(t)} = -\sum_{j=1}^{t} \underline{v}_j}$$

$sum^{(1)} = \underline{0}$

for $t = 1 \ldots T$

Let $\underline{w}^{(t)} = \left(\frac{1}{\lambda t}\right) sum^{(t)}$.

Choose $i \in [m]$

If $\left(y_i \langle \underline{w}^{(t)}, \underline{x}_i\rangle < 1\right)$ margin constraint is not satisfied.

$$sum^{(t+1)} = sum^{(t)} + y_i \underline{x}_i$$

else

$$sum^{(t+1)} = sum^{(t)}.$$