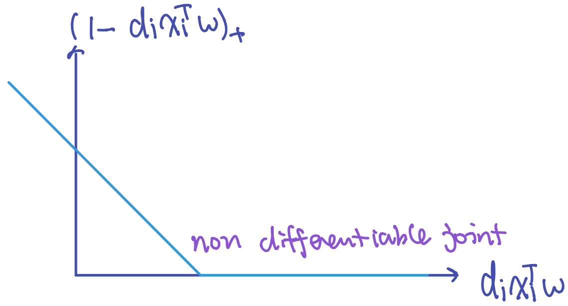
Support vector machines require iterative algorithms

$$\min_{w} \sum_{i=1}^{N} (1 - d_i x_i^T w)_+ + \lambda \|w\|_2^2$$

labels, features, hinge loss, regularization

No closed form solution.

but convex function $\Rightarrow$ gradient descent



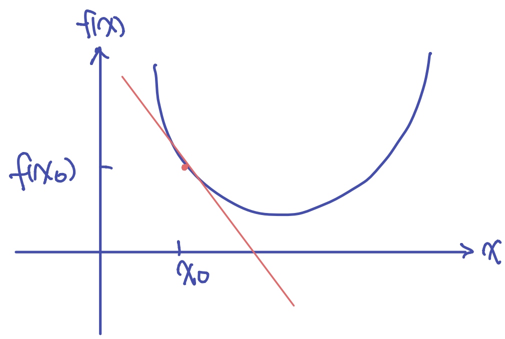
Problem: hinge loss non-differentiable

Subderivatives generalize derivatives

- Convex, but nondifferentiable $f(x)$

Derivatives –

$$d(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$



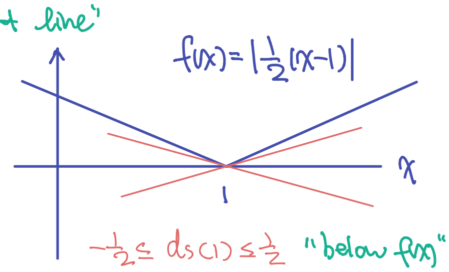
For a convex function:

$$f(x) \geq f(x_0) + d(x_0)(x - x_0) \quad \text{"above tangent line"}$$

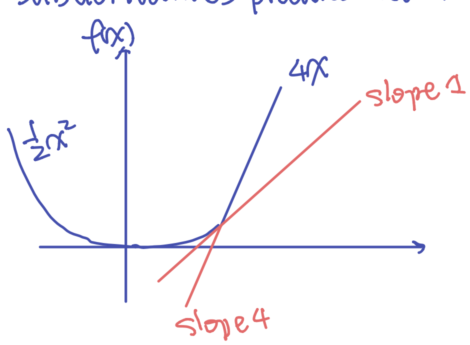
Subderivative (convex)

Any $d_s(x_0)$: $f(x) \geq f(x_0) + d_s(x_0)(x - x_0)$

$x < 1$: $d_s(x) = -\frac{1}{2}$; $x > 1$: $d_s(x) = \frac{1}{2}$



Subderivatives produce "reasonable" downhill directions



Example: $f(x) = \begin{cases} \frac{1}{2}x^2, & x < 1 \\ 4x, & x > 1 \end{cases}$ convex

Subderivative

$$d_s(x) = \begin{cases} x, & x < 1 \\ 4 & x > 1 \\ [1, 4] & x = 1 \end{cases}$$

Subgradients generalize gradients

- Convex, non differentiable $l(x)$

Gradients –

$$l(w) \geq l(w_0) + (w - w_0)^T v(w_0) \quad , \; v(w) = \nabla_w l(w)$$

"above tangent plane" $\left(\sum_{i=1}^{M} (w_i - w_{0i}) \frac{d}{dw_i} l(w) \right)$

Subgradients –

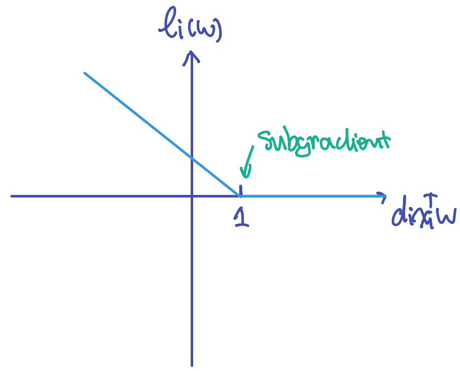
Any $v(w)$: $l(w) \geq l(w_0) + (w - w_0)^T v(w_0)$

Gradient descent optimization : replace gradient with subgradient

Gradient descent for SVMs

$$l(w) = \sum_{i=1}^{N} (1 - d_i x_i^T w_i)_+ \longrightarrow \text{subgradient}$$

$$l_i(w) = (1 - d_i x_i^T w)_+ = \begin{cases} 1 - d_i x_i^T w & d_i x_i^T w < 1 \\ 0 \; , & d_i x_i^T w > 1 \end{cases}$$



Subgradient

$$v_i(w) = \begin{cases} -d_i x_i & d_i x_i^T w < 1 \\ 0 & d_i x_i^T w \geq 0 \end{cases} = -d_i x_i I_{\{d_i x_i^T w < 1\}}$$

Indicator function

Cost $f(w) = l(w) + \lambda \|w\|_2^2$

$$\Rightarrow \nabla f(w)\big|_{w^{(k)}} = \sum_{i=1}^{N} \left(-d_i x_i I_{\{d_i x_i^T w^{(k)} < 1\}} \right) + 2\lambda w^{(k)}$$

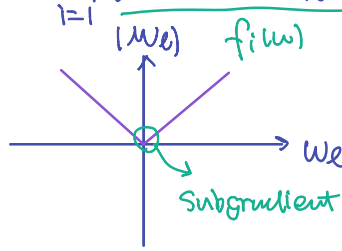
Gradient descent

$$w^{(k+1)} = w^{(k)} - \tau \nabla f(w)\big|_{w^{(k)}}$$

Example: Gradient Descent for LASSO

$$f(w) = \sum_{i=1}^{N} (d_i - x_i^T w)^2 + \lambda \|w\|_1 = \sum_{i=1}^{N} \underbrace{\left\{ (d_i - x_i^T w)^2 + \frac{\lambda}{N} \|w\|_1 \right\}}_{f_i(w)}$$

Consider $\nabla_w = \sum_{l=1}^{M} |w_l|$



$$\frac{d}{dw_l}|w_l| = \begin{cases} \text{sign}(w_l) & w_l \neq 0 \\ [-1, 1] & w_l = 0 \end{cases}$$

"0" popular

Write $\nabla_w \|w\|_1 = \text{sign}(w)$

$$\nabla_w f_i(w) = -2(d_i - x_i^T w) x_i + \frac{\lambda}{N} \text{sign}(w)$$

$$w^{(k+1)} = \tau(d_{i_k} - x_{i_k}^T w^{(k)}) x_{i_k} - \frac{\lambda \tau}{2N} \text{sign}(w^{(k)})$$