Gradient Descent for Support Vector Machines and Subgradients

Objectives

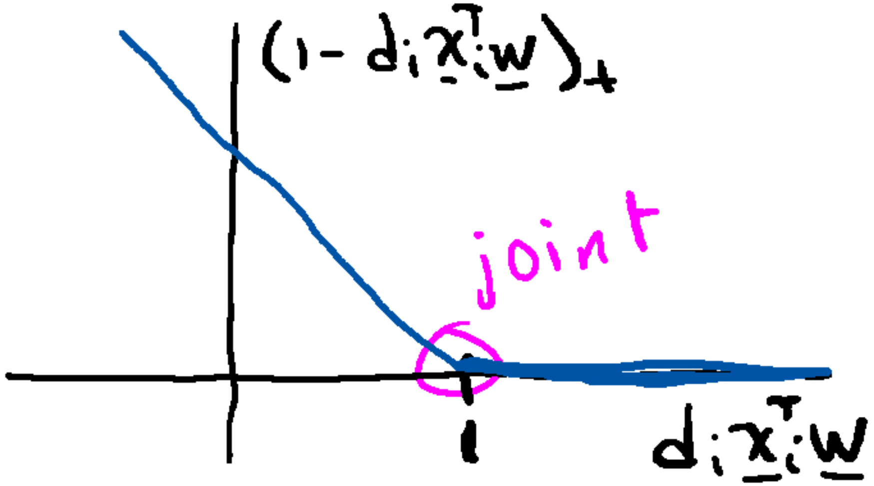
- develop a gradient descent algorithm for SVMs

- introduce subgradients for convex but nondifferentiable cost functions

Support vector machines require iterative algorithms

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

labels features hinge loss regularization



No closed form solution

Convex

$\Rightarrow$ gradient descent

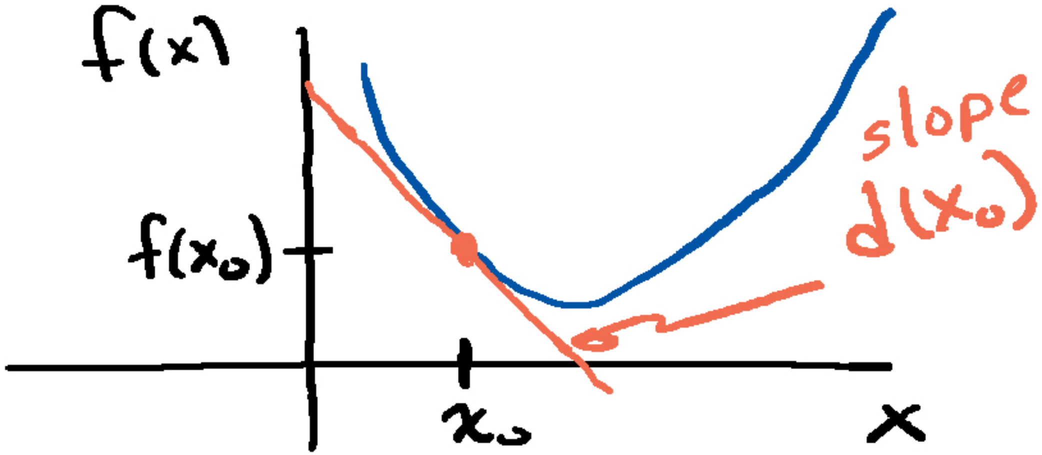
Problem: hinge loss not differentiable

Subderivatives generalize derivatives

- Convex, but nondifferentiable $f(x)$

Derivatives –

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



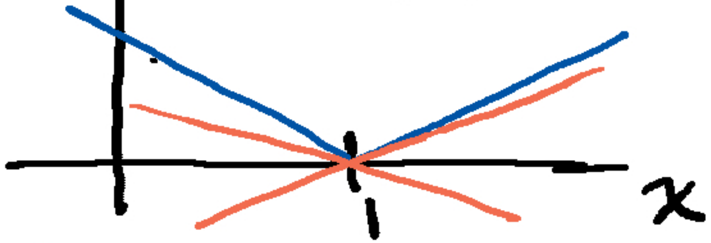
Convex:

$$f(x) \geq f(x_0) + d(x_0)(x - x_0)$$

"above tangent line"

Subderivative (convex)

$f(x) = |\frac{1}{2}(x-1)|$

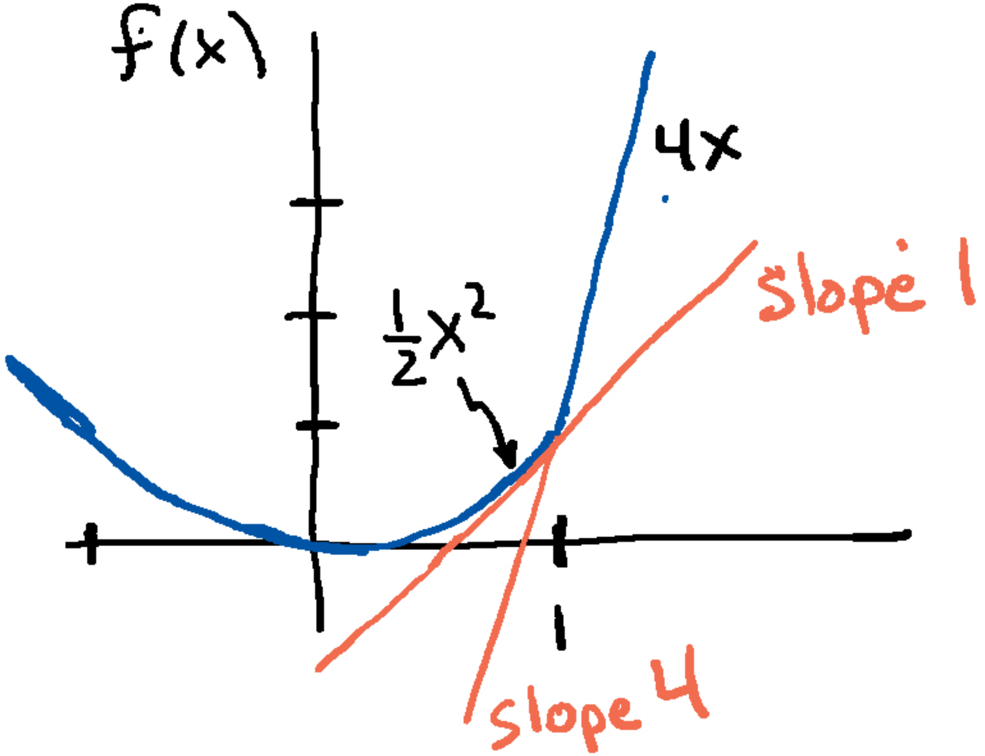


Any $d_s(x_0)$: $f(x) \geq f(x_0) + d_s(x_0)(x - x_0)$

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

$-\frac{1}{2} \leq d_s(1) \leq \frac{1}{2}$ "below $f(x)$"

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, nondifferentiable $\ell(\underline{w})$

Gradients -

$$\ell(\underline{w}) \geq \ell(\underline{w}_0) + (\underline{w}-\underline{w}_0)^T \underline{v}(\underline{w}_0) \qquad \underline{v}(\underline{w}) = \nabla_{\underline{w}} \ell(\underline{w})$$

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

Subgradients -

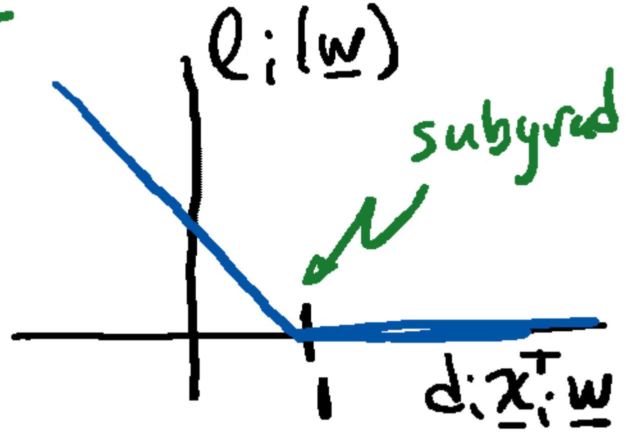
Any $\underline{v}(\underline{w})$: $\ell(\underline{w}) \geq \ell(\underline{w}_0) + (\underline{w}-\underline{w}_0)^T \underline{v}(\underline{w}_0)$

Gradient descent optimization: replace gradient with subgradient

Gradient descent for SVMs

$$\ell(\underline{w}) = \sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+ \rightarrow \text{subgradient}$$

$$\ell_i(\underline{w}) = (1 - d_i \underline{x}_i^T \underline{w})_+ = \begin{cases} 1 - d_i \underline{x}_i^T \underline{w} & d_i \underline{x}_i^T \underline{w} < 1 \\ 0 & d_i \underline{x}_i^T \underline{w} > 1 \end{cases}$$



Subgradient

$$\underline{v}_i(\underline{w}) = \begin{cases} -d_i \underline{x}_i & d_i \underline{x}_i^T \underline{w} < 1 \\ 0 & d_i \underline{x}_i^T \underline{w} \geq 1 \end{cases} = -d_i \underline{x}_i \, I_{\{d_i \underline{x}_i^T \underline{w} < 1\}}$$

indicator function

Cost $f(\underline{w}) = \ell(\underline{w}) + \lambda \|\underline{w}\|_2^2$

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

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