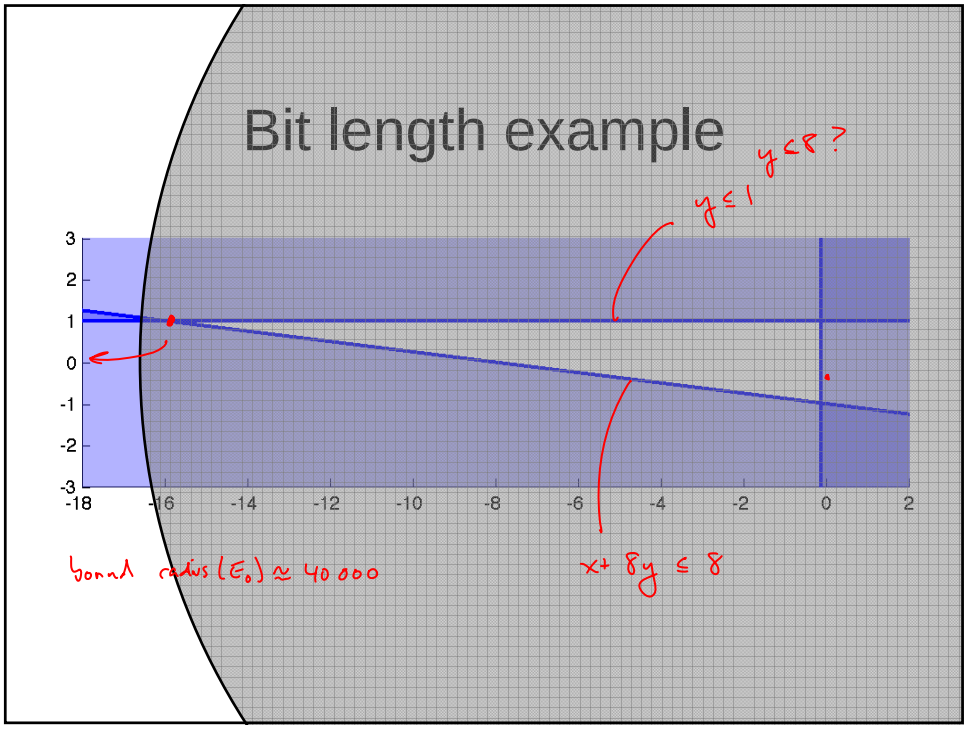
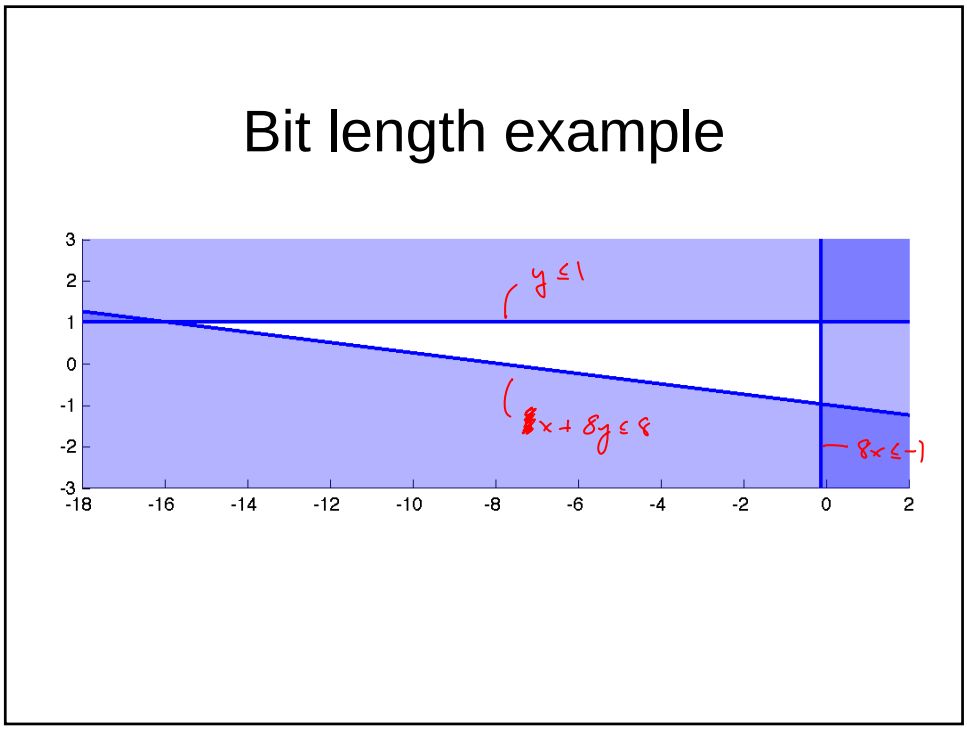
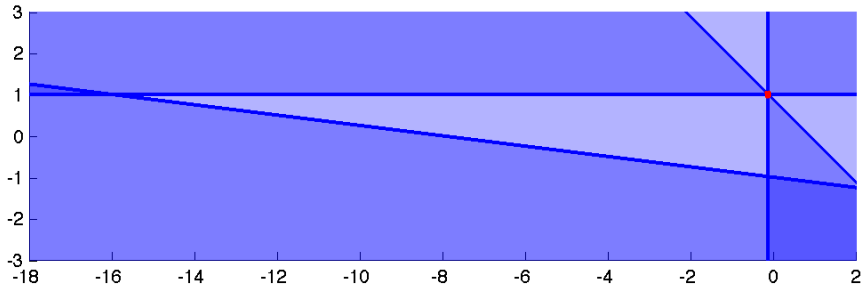
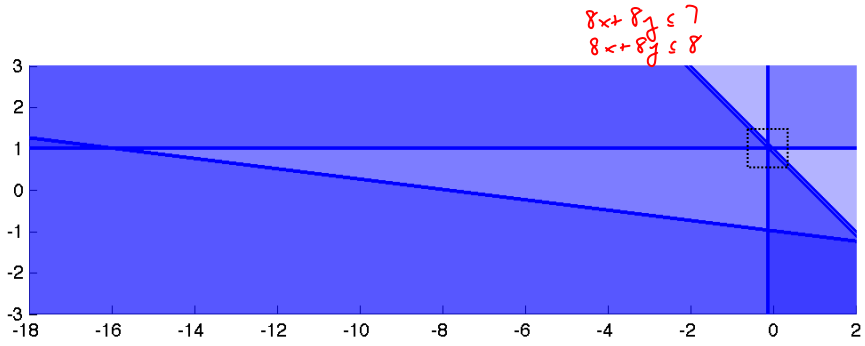

Bit length example
y ≤ 8 ?
y ≤ 1
3
2
1
0
-1
-2
-3
-18
-16
-14
-12
-10
-8
-6
-4
-2
0
2
bound radius (E₀) ≈ 40 000
x + 8y ≤ 8


Bit length example
3
2
1
0
-1
-2
-3
-18
-16
-14
-12
-10
-8
-6
-4
-2
0
2
y ≤ 1
x + 8y ≤ 8
8x ≤ -1

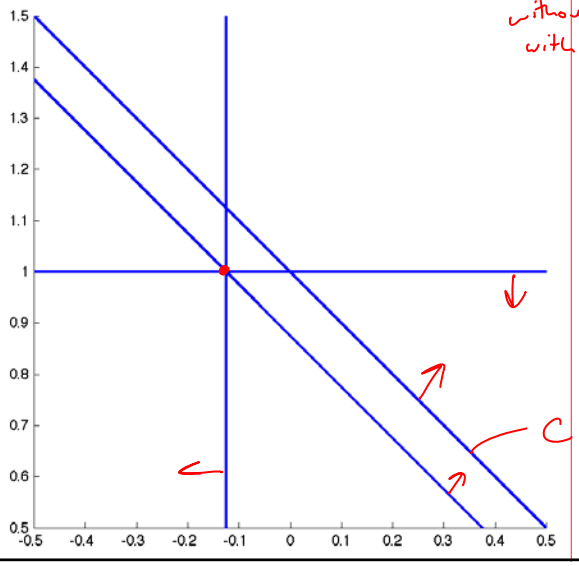
Bit length example



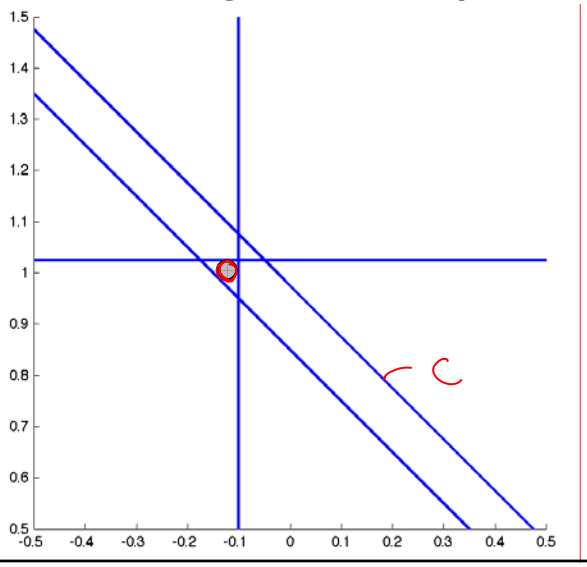
Bit length example



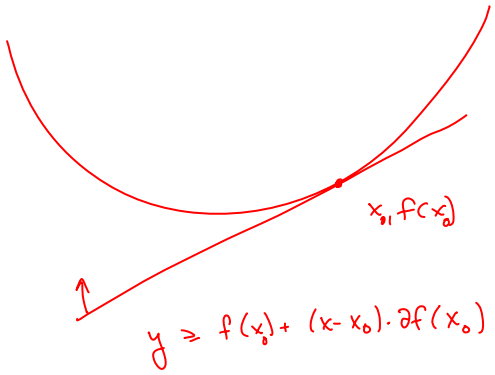
Bit length example



Bit length example

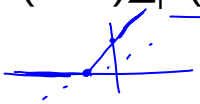


What's a subgradient?



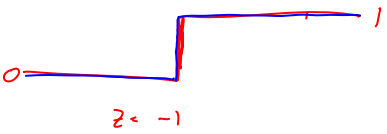
Subgradients for SVMs

- $\min_w L(w) = ||w||^2 + (C/m)\sum_i h(-y_i x_i^T w)$
- $h(z) = \max\{0, 1+z\}$



- Subgradient of $h(z)$:

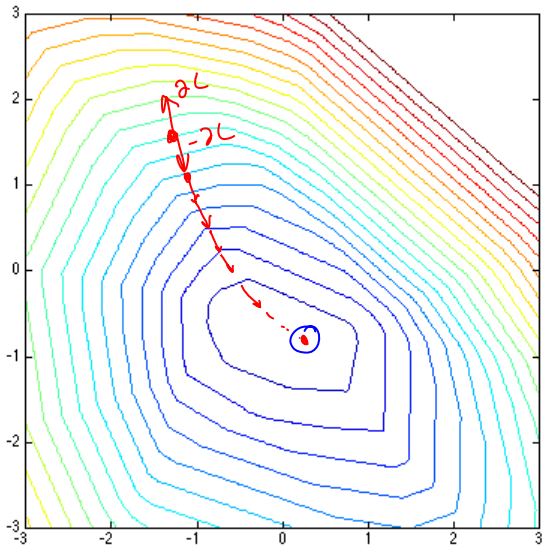
$$\partial(h(z)) = \begin{cases} 0 & z < -1 \\ 1 & z > -1 \\ [0,1] & z = -1 \end{cases}$$



- Subgradient of $L(w)$ wrt w:

$$\partial L(w) = 2w + \frac{C}{m} \sum_i \partial h(-y_i x_i^T w) \cdot (-y_i x_i)$$

Subgradient descent



Subgradient descent

Start w/ x_0

- While not tired:

  $g_t =$ (estimate of) $\partial f(x_t)$

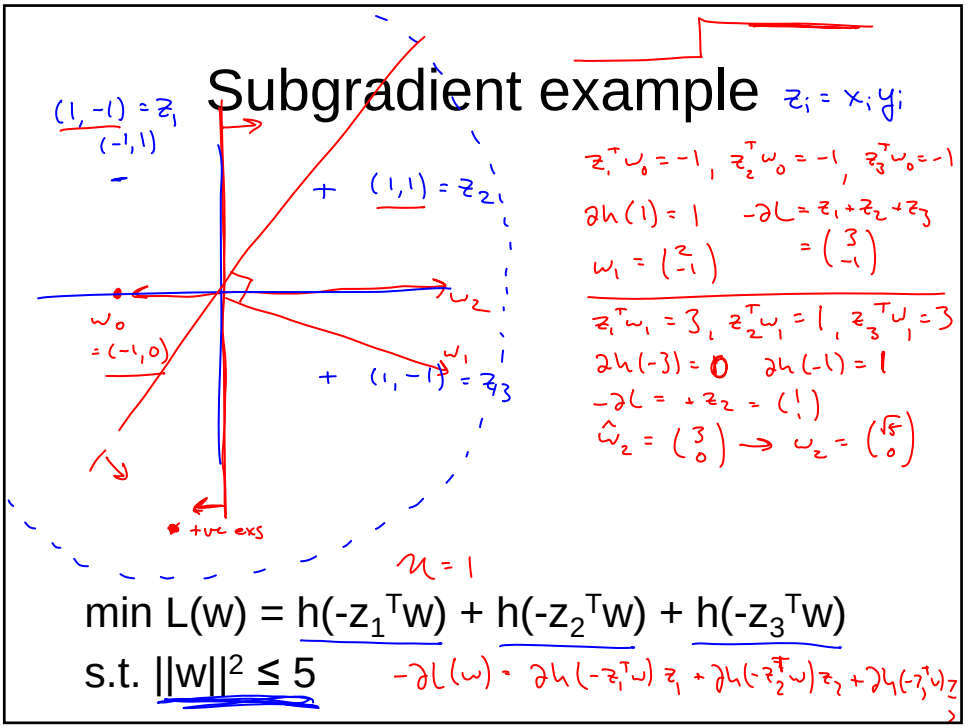
  $x_{t+1} = x_t - \eta_t g_t$

  $x_{t+1} := \Pi_F x_{t+1}$

  η_t = learning rate

  Π_F: projection onto feasible region F

Subgradient example



$z_i = x_i y_i$

$z_1^T w_0 = -1,\ z_2^T w_0 = -1,\ z_3^T w_0 = -1$

$\partial h(1) = 1 \quad -\partial L = z_1 + z_2 + z_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$

$w_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$

$z_1^T w_1 = 3,\ z_2^T w_1 = 1,\ z_3^T w_1 = 3$

$\partial h(-3) = 0 \quad \partial h(-1) = 1$

$-\partial L = + z_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$\hat{w}_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \rightarrow w_2 = \begin{pmatrix} \sqrt{5} \\ 0 \end{pmatrix}$

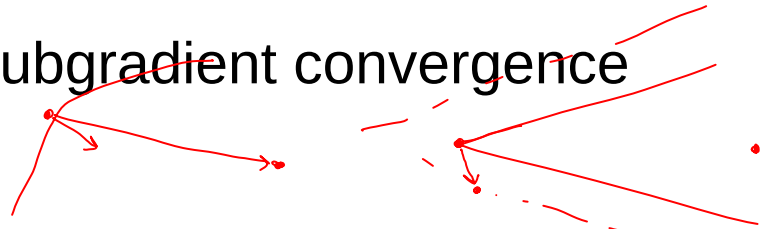
$\eta = 1$

$\min L(w) = h(-z_1^T w) + h(-z_2^T w) + h(-z_3^T w)$

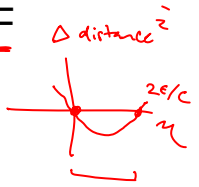
s.t. $\|w\|^2 \le 5$

$-\partial L(w) = \partial h(-z_1^T w) z_1 + \partial h(-z_2^T w) z_2 + \partial h(-z_3^T w) z_3$

Subgradient convergence



- Suppose $\|\partial L(x)\|^2 \le C$ for all x in F
- Suppose $L(x_t) \ge L(x^*) + \varepsilon$ → optimum



$L(x^*) \ge L(x_t) + (x^* - x_t) \cdot \partial L(x_t)$

$(x_* - x_t)(-\partial L(x_t)) \ge L(x_t) - L(x_*) \ge \epsilon$

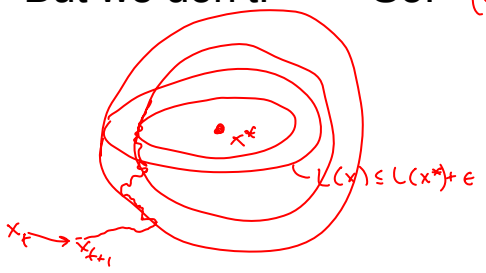
$\|x^* - x_{t+1}\|^2 = \|x^* - (x_t - \eta \partial L(x_t))\|^2$

$= \|x^* - x_t\|^2 + \eta^2 \|\partial L(x_t)\|^2 + 2\eta (x^* - x_t) \partial L(x_t)$

$< \|x^* - x_t\|^2 + \eta^2 C - 2\eta\epsilon$

Setting step size

- If we knew ε, could set good step size η
- But we don't! So: $\eta_t \to 0$ as $t \to \infty$

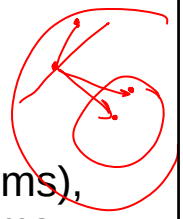


if $\eta_t \le \epsilon / C$ → bounds $\|\partial L\|$

$\sum_{t=1}^{\infty} \eta_t = \infty$

- Typical choices: $\eta_t = \eta$ $\eta_t = O(1/t)$ $\eta_t = O(1/\sqrt{t})$

Stochastic subgradient



- In SVM (and many other ML problems), L(w) contains big sum of simple terms

  $\min_w L(w) = ||w||^2 + (C/m)\sum_i h(-y_i x_i^T w)$

  $\partial L(w) = 2w - \frac{C}{m}\sum_i \partial h(-y_i x_i \cdot w)(y_i x_i)$ (norm of term i) $\le \|x_i\|$
- Approximate sum by sampling terms

  $\|\partial_i\| \le C\|x_i\|$ $\partial_i = C\,\partial h(-y_i x_i \cdot w) y_i x_i$ $\partial_S = 2w - \frac{C}{|S|}\sum_{i \in S} \partial_i$ $E(\partial_S) = \partial L(w)$

  $E(-\partial_S^T(x - x^*)) = -\partial L(w) \cdot (x - x^*)$

  S random, $|S| = k$: $\mathrm{Var}(\eta \partial_S) \le \eta^2 \frac{C^2 \|x\|^2}{k} = O(\eta^2)$

When do we stop?

- Feasible region diameter ||F||

$f(x_t) \geq f(x^*) \geq f(x_t) + (x^* - x_t) \cdot \partial f(x_t) \geq f(x_t) - \|x^* - x_t\| \|\partial f(x_t)\|$

$\geq f(x_t) - \|F\| \|\partial f(x_t)\|$

- Typical ML generalization bound:

$\forall w$ $E(L(\text{new ex}, w)) \leq L(\text{train}, w) + \text{stuff}$

$w^* = \text{opt}$

$L(\text{train}, w') \leq L(\text{train}, w^*) + O(\text{stuff})$

$E(L(\text{new ex}, w')) \leq L(\text{train}, w') + \text{stuff}$

$\leq L(\text{train}, w^*) + O(\text{stuff}) + \text{stuff}$ → $O(\text{stuff})$