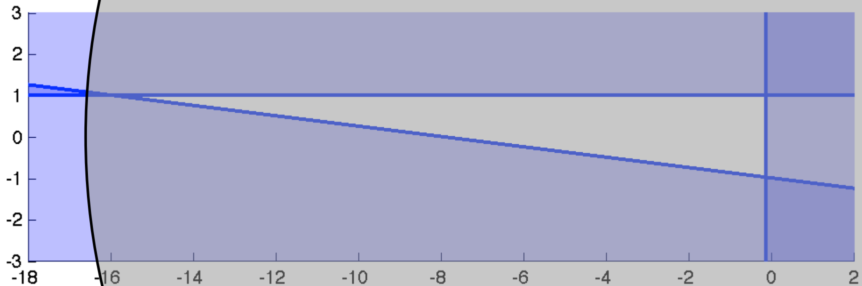
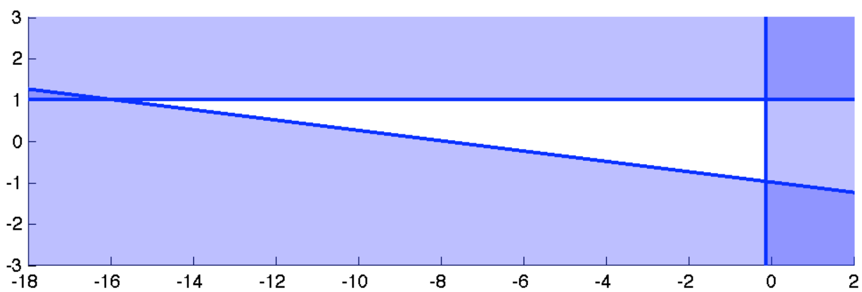
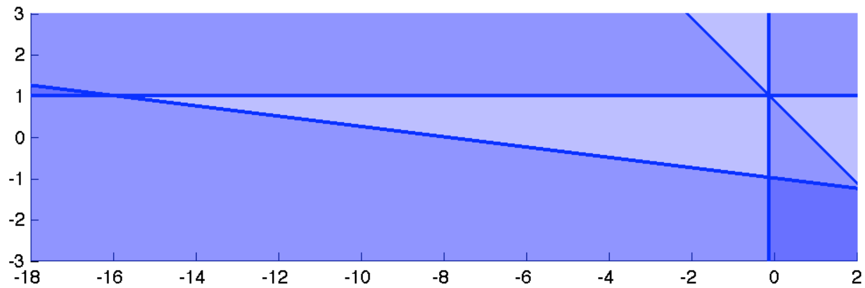
Bit length example



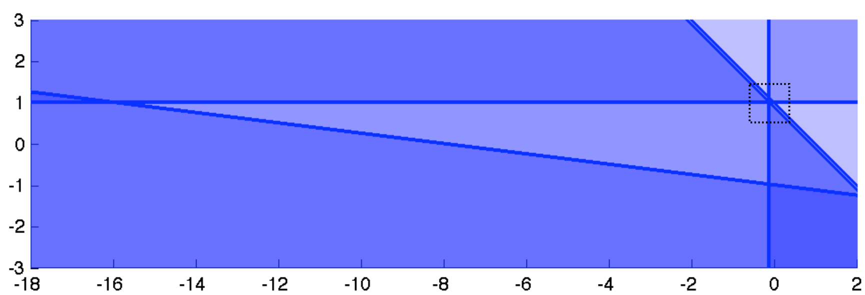
Bit length example



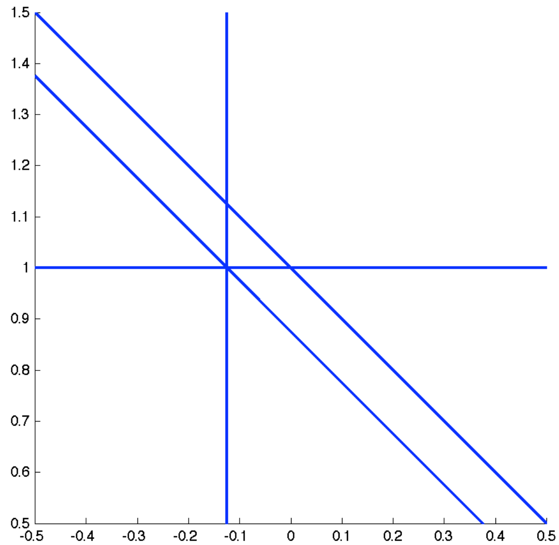
Bit length example



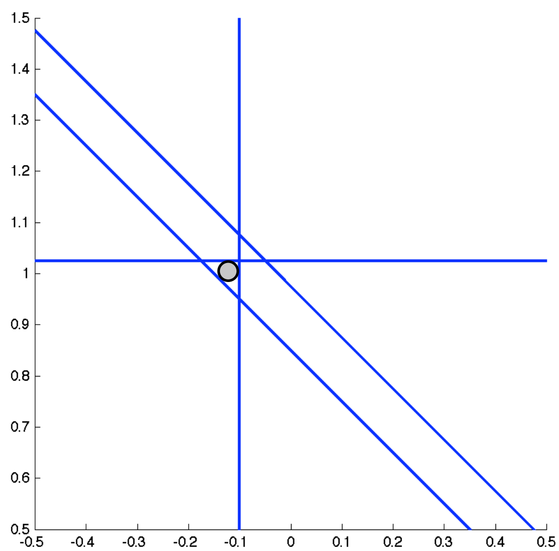
Bit length example



Bit length example



Bit length example

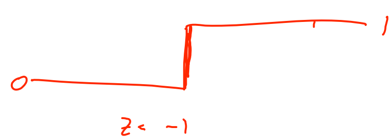


What's a subgradient?

Subgradients for SVMs

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

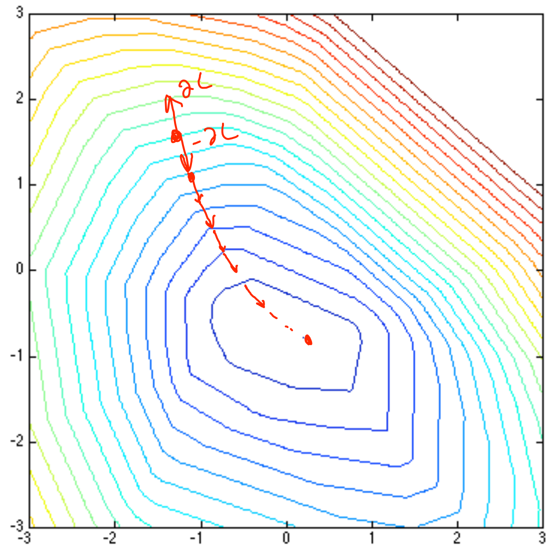
$$\partial(h(z)) = \begin{cases} 0 & z < -1 \\ 1 & z > -1 \\ [0,1] & z = -1 \end{cases}$$



- 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) $\partial f(x_t)$

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

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

  ↑ projection onto feasible region F

η_t = learning rate

Subgradient example

$\min L(w) = h(-z_1^T w) + h(-z_2^T w) + h(-z_3^T w)$
s.t. $||w||^2 \leq 5$

Subgradient convergence

- Suppose $||\partial L(x)||^2 \leq C$ for all x in F
- Suppose $L(x_t) \geq L(x^*) + \varepsilon$

Setting step size

- If we knew ε, could set good step size η
- But we don't! So:

- Typical choices:

Stochastic subgradient

- In SVM (and many other ML problems), L(w) contains big sum of simple terms
  $\min_w L(w) = ||w||^2 + (C/m)\Sigma_i h(-y_i x_i^T w)$
  $\partial L(w) =$
- Approximate sum by sampling terms
  $\partial_i =$ $\partial_S =$
  $E(-\partial_S^T(x-x^*)) =$
  S random, $|S| = k$: $Var(\eta\partial_S) \leq$

When do we stop?

- Feasible region diameter ||F||
  $\geq f(x^*) \geq$

- Typical ML generalization bound:
  E(L(new ex, w)) ≤ L(train, w) +