Bit length example

Bit length example
What’s a subgradient?

Subgradients for SVMs

• $\min_w L(w) = ||w||^2 + (C/m) \sum_i h(-y_ix_i^Tw)$
• $h(z) = \max \{0, 1+z\}$
• Subgradient of $h(z)$:

$$\partial h(z) = \begin{cases} 0 & z \leq -1 \\ 1 & z > -1 \\ [0,1] & z = -1 \\ 0 & 2z = -1 \end{cases}$$

• Subgradient of $L(w)$ wrt $w$:

$$\partial L(w) = 2w + \frac{C}{m} \sum_i h(-y_ix_i^Tw), (-y_ix_i)$$
Subgradient descent

• While not tired:

\[ g_t = \text{(estimate } \partial f(x_t) \text{)} \]

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

\[ x_{t+1} \overset{\text{projection}}{=} \cap_{F} x_{t+1} \]

\[ \eta_t = \text{learning rate} \]

Start w/ \( x_0 \)
Subgradient example

\[ \min L(w) = h(-z_1^Tw) + h(-z_2^Tw) + h(-z_3^Tw) \]
\[ \text{s.t. } \|w\|^2 \leq 5 \]

Subgradient convergence

- Suppose \( \|\partial L(x)\|^2 \leq C \) for all x in F
- Suppose \( L(x_t) \geq L(x^*) + \epsilon \)
Setting step size

• If we knew $\varepsilon$, could set good step size $\eta$
• But we don’t! So:

• Typical choices:

Stochastic subgradient

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

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

$$\partial L(w) =$$

• Approximate sum by sampling terms

$$\partial_i =$$

$$\partial_S =$$

$$\mathbb{E}(-\partial_S^T (x-x^*)) =$$

$S$ random, $|S| = k$: $\text{Var}(\eta \partial_S) \leq$
When do we stop?

• Feasible region diameter $||F|| ≥ f(x^*) ≥$

• Typical ML generalization bound:
  $E(L\text{(new ex, w)}) ≤ L\text{(train, w)} +$