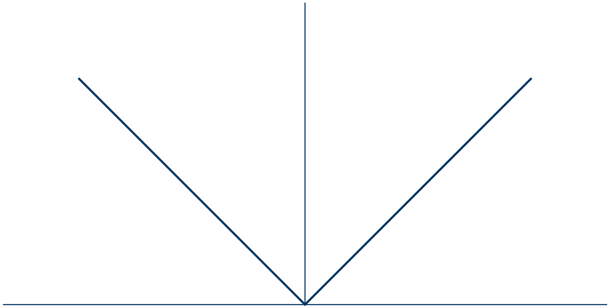
Machine Learning: Optimization 3

Hao Tang

February 3, 2026

Subgradients for absolute values

$|x|$



Subgradient

- A subgradient at x is a vector g that satisfies

$$f(y) \geq f(x) + g^\top (y - x) \tag{1}$$

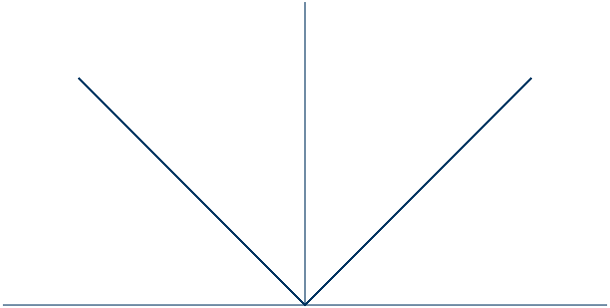
  for any y.

- In other words, a subgradient defines a supporting hyperplane.

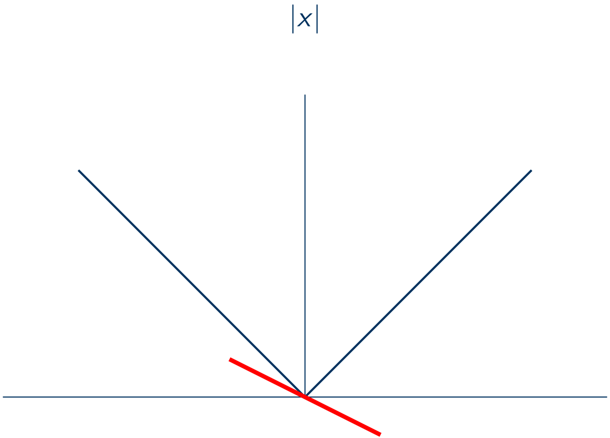
- In fact, any supporting hyperplane gives a subgradient, so a subgradient, unlike the gradient, might not be unique.

Subgradients for absolute values

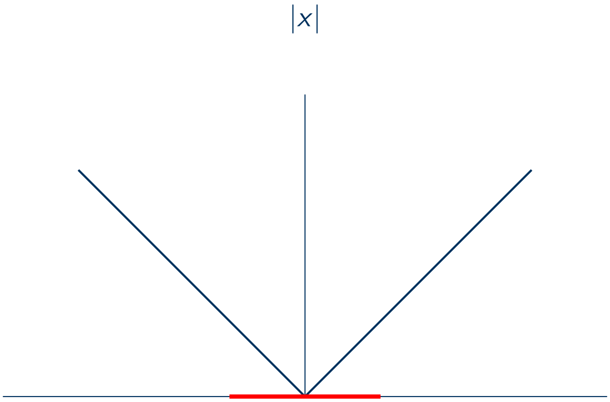
$|x|$



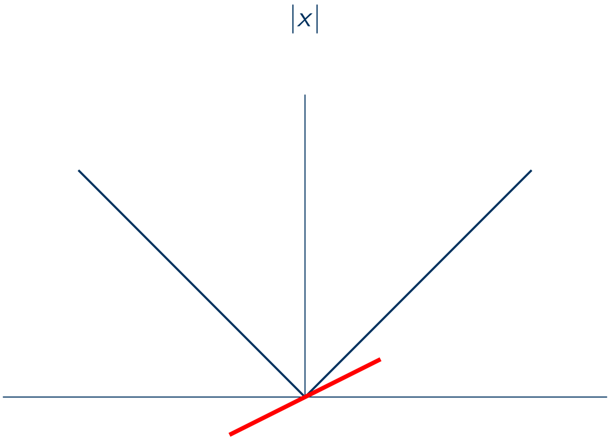
Subgradients for absolute values



Subgradients for absolute values



Subgradients for absolute values



Hinge loss

- The hinge loss is defined as

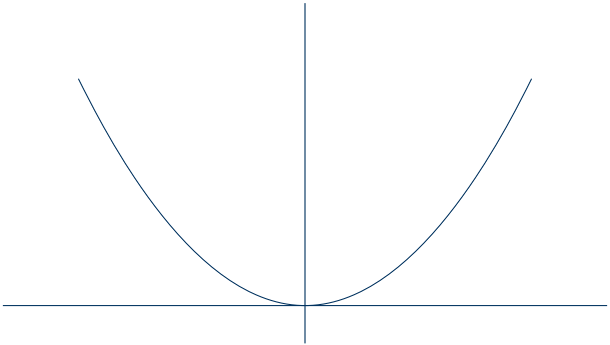
$$\ell_{\text{hinge}}(w; x, y) = \max(0, 1 - yw^\top x). \tag{2}$$

- Just like the absolute value, the hinge loss is continuous and convex, but it is not differentiable.

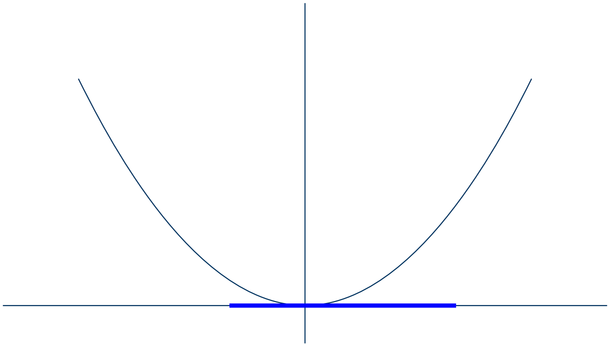
$$\nabla_w \ell = \begin{cases} 0 & \text{if } yw^\top x \geq 1 \\ -yx & \text{if } yw^\top x < 1 \end{cases} \tag{3}$$

- When $yw^\top x = 1$, we can pick and choose any vector that supports the loss function from below as the subgradient. In fact, 0 and $-yx$ both work.

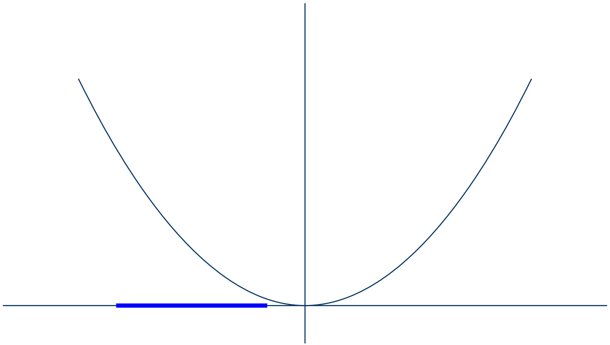
Constrained optimization



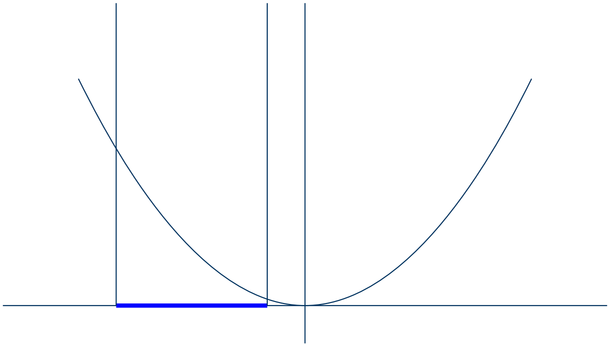
Constrained optimization



Constrained optimization



Setting up a barrier



An example problem with constraints

- The problem

$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & -2.5 \leq x \leq -0.5 \end{aligned} \tag{4}$$

  is an example of a contrained optimization problem.

- The inequality $-2.5 \leq x \leq -0.5$ is called a constraint.

- Solutions that satisfy the constraints are called **feasible** solutions.

Setting up a barrier

- The problem

$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & -2.5 \leq x \leq -0.5 \end{aligned} \tag{5}$$

is equivalent to

$$\min_x x^2 + V_-(x) \tag{6}$$

if

$$V_-(x) = \begin{cases} 0 & \text{if } -2.5 \leq x \leq -0.5 \\ \infty & \text{otherwise} \end{cases} \tag{7}$$

Another example problem with constraints

- The problem

$$\begin{aligned} \min_{w} \quad & L(w) \\ \text{s.t.} \quad & \|w\|^2 \leq 1 \end{aligned} \tag{8}$$

  is an example of a contrained optimization problem.

- The inequality $\|w\|^2 \leq 1$ is called a constraint.

Setting up a barrier

- We can write the optimization problem as

$$\min_{w} \qquad L(w) + V_{-}(\|w\|_2^2 - 1), \tag{9}$$

where

$$V_{-}(s) = \begin{cases} 0 & \text{if } s \leq 0 \\ \infty & \text{if } s > 0 \end{cases}. \tag{10}$$

Setting up a barrier

- Setting up the barrier moves the constraints to the objective function.
- This technique reduces the problem of constrained optimization back to unconstrained optimization.
- This does not change anything; both problems are equally hard (or easy) to solve.

Soften the constraints

- We can linearize the barrier and turn

$$\min_{w} \quad L(w) + V_{-}(\|w\|_2^2 - 1) \tag{11}$$

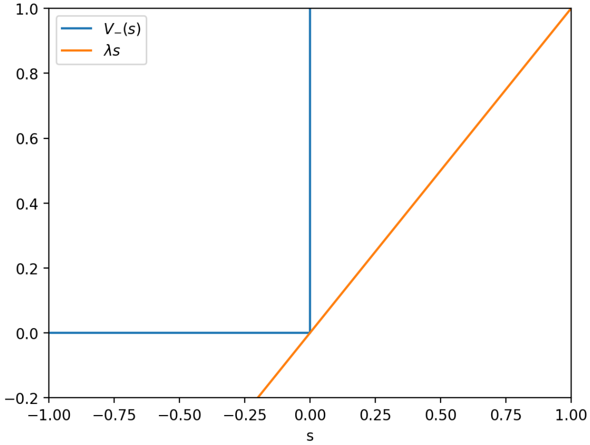
  into

$$\min_{w} \quad L(w) + \lambda(\|w\|_2^2 - 1), \tag{12}$$

  for some $\lambda \geq 0$.

- Note that $\lambda s \leq V_{-}(s)$ for all s.
- In other words, the linearized objective value is always lower than the one with the barrier.

Soften the constraints



Soften the constraints

- We start with this constrained optimization problem

$$\begin{aligned} \min_{w} \quad & L(w) \\ \text{s.t.} \quad & \|w\|^2 \leq 1 \end{aligned} \tag{13}$$

- We end up with the problem

$$\min_{w} \quad L(w) + \lambda(\|w\|_2^2 - 1) \tag{14}$$

which is just an ordinary unconstrained optimization, and we know how to solve it.

Lagrangian

- In general, if we have an optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & h(x) \leq 0 \end{aligned} \tag{15}$$

the **Lagrangian** is defined as

$$F(x, \lambda) = f(x) + \lambda h(x). \tag{16}$$

- The value $\lambda \geq 0$ is called the Lagrange multiplier.

A unigram model

Row, row, row your boat, gently down the stream
Merrily, merrily, merrily, merrily, life is but a dream

A unigram model

Row, row, row your boat, gently down the stream
Merrily, merrily, merrily, merrily, life is but a dream

- There are 18 words.
- Intuitively,

$$p(\text{row}) = \frac{3}{18} \qquad p(\text{merrily}) = \frac{4}{18} \qquad p(\text{is}) = \frac{1}{18} \tag{17}$$

A unigram model

- There are 13 unique words.
- We refer to the set of unique words $V = \{\text{row}, \text{your}, \text{boat}, \text{gently}, \text{down}, \text{the}, \text{stream}, \text{merrily}, \text{life}, \text{is}, \text{but}, \text{a}, \text{dream}\}$ as the vocabulary.
- The goal is to estimate the probability of each word, i.e., figuring out what the β's are in the table.

v	row	your	boat	...
β_v	β_{row}	β_{your}	β_{boat}	...

A unigram model

- We assign each word v a probability β_v.
- Since β is a probability vector, we have the constraint

$$\sum_{v \in V} \beta_v = 1. \tag{18}$$

- The probability of a word is

$$p(w) = \prod_{v \in V} \beta_v^{\mathbb{1}_{v=w}}. \tag{19}$$

A unigram model

- We assume that each word is independent of others.
- This assumption is obviously wrong, but can go really far.
- The likelihood of β given the data is

$$\log p(w_1, \ldots, w_N) = \log \prod_{i=1}^{N} p(w_i) = \log \prod_{i=1}^{N} \prod_{v \in V} \beta_v^{\mathbb{1}_{v=w_i}}. \tag{20}$$

A unigram model

- We arrive at the optimization problem

$$
\begin{aligned}
\min_{\beta} \quad & -\sum_{i=1}^{N} \sum_{v \in V} \mathbb{1}_{v=w_i} \log \beta_v \\
\text{s.t.} \quad & \sum_{v \in V} \beta_v = 1
\end{aligned}
\tag{21}
$$

- Its Lagrangian is

$$
F = -\sum_{i=1}^{N} \sum_{v \in V} \mathbb{1}_{v=w_i} \log \beta_v + \lambda \left(\sum_{v \in V} \beta_v - 1 \right). \tag{22}
$$

A unigram model

- Solving the optimality condition gives

$$\frac{\partial F}{\partial \beta_k} = \sum_{i=1}^{N} \mathbb{1}_{k=w_i} \frac{1}{\beta_k} - \lambda = 0 \implies \beta_k = \frac{1}{\lambda} \sum_{i=1}^{N} \mathbb{1}_{k=w_i}. \tag{23}$$

A unigram model

$$\sum_{v \in V} \beta_v = \sum_{v \in V} \frac{1}{\lambda} \sum_{i=1}^{N} \mathbb{1}_{v=w_i} = 1 \implies \lambda = \sum_{v \in V} \sum_{i=1}^{N} \mathbb{1}_{v=w_i} = N \tag{24}$$

$$\beta_k = \frac{\sum_{i=1}^{N} \mathbb{1}_{k=w_i}}{\sum_{v \in V} \sum_{i=1}^{N} \mathbb{1}_{v=w_i}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{k=w_i} \tag{25}$$

Solving the Lagrangian

- Find $\hat{x} = \text{argmin}_x[f(x) + \lambda h(x)]$ for any λ.
- For example, one approach to finding $\hat{x}$ is to solve

$$\nabla_x[f(x) + \lambda h(x)] = 0 \tag{26}$$

- Find $\hat{\lambda}$ such that $\lambda h(\hat{x}) = 0$.
- The pair $\hat{x}$ and $\hat{\lambda}$ gives a feasible and optimal solution (if they exist).

Why solving the Lagrangian works

- Suppose $\hat{x} = \text{argmin}_x[f(x) + \lambda h(x)]$ and $x^* = \text{argmin}_{x:h(x)\leq 0} f(x)$.

$$f(\hat{x}) + \lambda h(\hat{x}) \leq f(x^*) + \lambda f(x^*) \leq f(x^*) \tag{27}$$

- If $\lambda h(\hat{x}) = 0$, then $\hat{x}$ is an optimial solution.

- If $\hat{x}$ is an optimial solution, then $\lambda h(\hat{x}) = 0$.

Complementary slackness

- There are two cases where $\lambda h(x)$ can be 0.
  1. One is that $\lambda = 0$ and $h(x) < 0$.

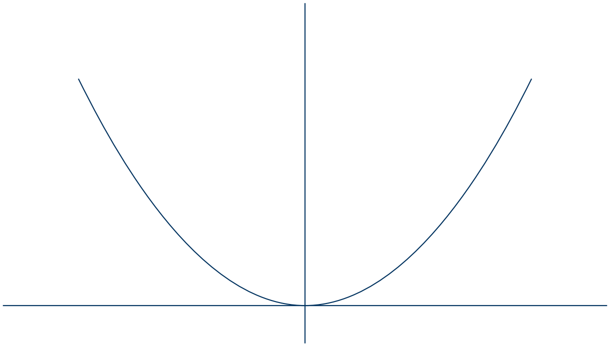
     In this case, the optimal solution is within the constraint set.

  2. The other is that $\lambda > 0$ and $h(x) = 0$.

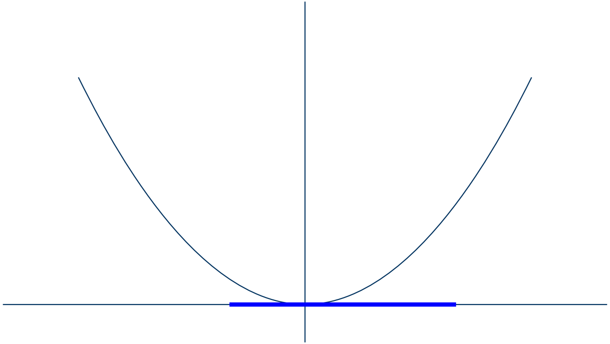
     In this case, the optimal solution is on the boundary of the constraint set.

- The condition $\lambda h(x) = 0$ is so important that it has a name called complementary slackness.

Complementary slackness



Complementary slackness



Complementary slackness