Subgradient Descent

He He

Slides based on Lecture 3c from David Rosenberg's course material.

CDS, NYU

Feb 23, 2021

Subgradient Descent

He He

CDS, NYU

Feb 23, 2021

SVM Optimization Problem (no intercept)

- SVM objective function:

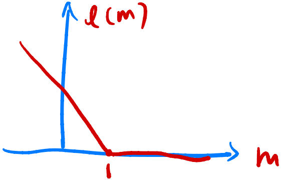
$$J(w) = \frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i w^T x_i\right) + \lambda \|w\|^2.$$

- Not differentiable... but let's think about gradient descent anyway.
- Hinge loss: $\ell(m) = \max(0, 1-m)$

$$\begin{aligned}
\nabla_w J(w) &= \nabla_w \left(\frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i w^T x_i\right) + \lambda \|w\|^2 \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \nabla_w \ell\left(y_i w^T x_i\right) + 2\lambda w
\end{aligned}$$

"Gradient" of SVM Objective

- Derivative of hinge loss $\ell(m) = \max(0, 1 - m)$:



$$\ell'(m) = \begin{cases} 0 & m > 1 \\ -1 & m < 1 \\ \text{undefined} & m = 1 \end{cases}$$

- By chain rule, we have

$$\begin{aligned} \nabla_w \ell\left(y_i w^T x_i\right) &= \ell'\left(\underbrace{y_i w^T x_i}_{m}\right) y_i x_i \\ &= \begin{cases} 0 & y_i w^T x_i > 1 \\ -y_i x_i & y_i w^T x_i < 1 \\ \text{undefined} & y_i w^T x_i = 1 \end{cases} \end{aligned}$$

"Gradient" of SVM Objective

$$\nabla_w \ell\left(y_i w^T x_i\right) = \begin{cases} 0 & y_i w^T x_i > 1 \\ -y_i x_i & y_i w^T x_i < 1 \\ \text{undefined} & y_i w^T x_i = 1 \end{cases}$$

So

$$\begin{aligned} \nabla_w J(w) &= \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i w^T x_i\right) + \lambda \|w\|^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \nabla_w \ell\left(y_i w^T x_i\right) + 2\lambda w \\ &= \begin{cases} \frac{1}{n} \sum_{i: y_i w^T x_i < 1} \left(-y_i x_i\right) + 2\lambda w & \text{all } y_i w^T x_i \neq 1 \\ \text{undefined} & \text{otherwise} \end{cases} \end{aligned}$$

Gradient Descent on SVM Objective?

- The gradient of the SVM objective is

$$\nabla_w J(w) = \frac{1}{n} \sum_{i: y_i w^T x_i < 1} (-y_i x_i) + 2\lambda w$$

when $y_i w^T x_i \neq 1$ for all i, and otherwise is undefined.

Potential arguments for why we shouldn't care about the points of nondifferentiability:

- If we start with a random w, will we ever hit exactly $y_i w^T x_i = 1$?
- If we did, could we perturb the step size by ε to miss such a point?
- Does it even make sense to check $y_i w^T x_i = 1$ with floating point numbers?

However, would gradient descent work if the objective is not differentiable?

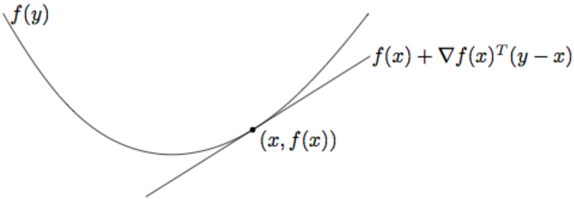
Subgradient

First-Order Condition for Convex, Differentiable Function

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable Then for any $x, y \in \mathbb{R}^d$

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x)$$

- The linear approximation to f at x is a global underestimator of f:



- This implies that if $\nabla f(x) = 0$ then x is a global minimizer of f.

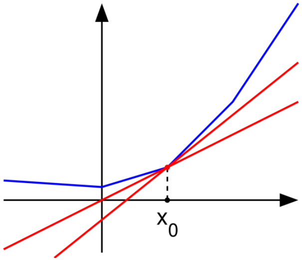
Figure from Boyd & Vandenberghe Fig. 3.2; Proof in Section 3.1.3

Subgradients

Definition

A vector $g \in \mathbf{R}^d$ is a **subgradient** of a *convex* function $f : \mathbf{R}^d \to \mathbf{R}$ at x if for all z,

$$f(z) \geqslant f(x) + g^T(z - x).$$



Blue is a graph of $f(x)$.
Each red line $x \mapsto f(x_0) + g^T(x - x_0)$ is a global lower bound on $f(x)$.

Properties

Definitions

- The set of all subgradients at x is called the **subdifferential**: $\partial f(x)$
- f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.

For convex functions:

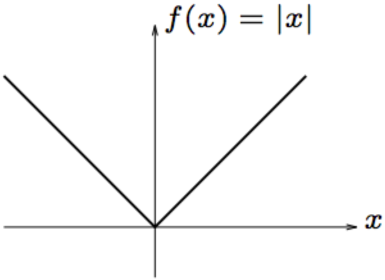
- f is differentiable at x iff $\partial f(x) = \{\nabla f(x)\}$.
- Subdifferential is always non-empty ($\partial f(x) = \emptyset \implies f$ is not convex)
- x is the global optimum iff $0 \in \partial f(x)$.

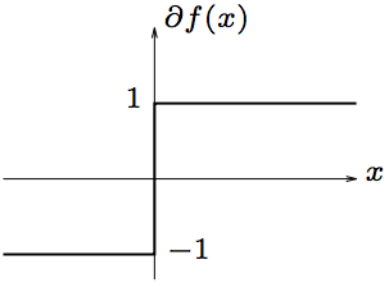
For non-convex functions:

- The subdifferential may be an empty set (no global underestimator).

Subdifferential of Absolute Value

- Consider $f(x) = |x|$



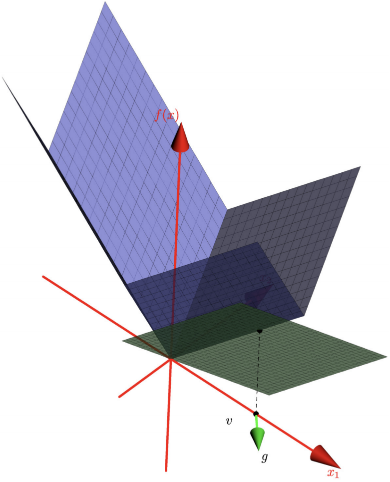


- Plot on right shows $\{(x, g) \mid x \in \mathbf{R},\ g \in \partial f(x)\}$

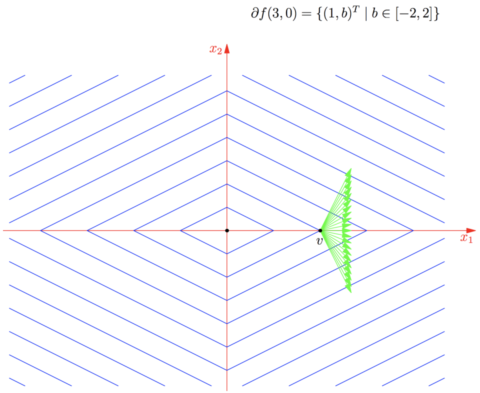
Boyd EE364b: Subgradients Slides

Subgradients of $f(x_1, x_2) = |x_1| + 2|x_2|$

- Let's find the subdifferential of $f(x_1, x_2) = |x_1| + 2|x_2|$ at $(3, 0)$.
- First coordinate of subgradient must be 1, from $|x_1|$ part (at $x_1 = 3$).
- Second coordinate of subgradient can be anything in $[-2, 2]$.
- So graph of $h(x_1, x_2) = f(3, 0) + g^T(x_1 - 3, x_2 - 0)$ is a global underestimate of $f(x_1, x_2)$, for any $g = (g_1, g_2)$, where $g_1 = 1$ and $g_2 \in [-2, 2]$.



Subdifferential on Contour Plot



Contour plot of $f(x_1,x_2) = |x_1| + 2|x_2|$, with set of subgradients at $(3,0)$. .

Plot courtesy of Brett Bernstein.

Basic Rules for Calculating Subdifferential

- Non-negative scaling: $\partial \alpha f(x) = \alpha \partial f(x)$ for $(\alpha > 0)$
- Summation: $\partial(f_1(x) + f_2(x)) = d_1 + d_2$ for any $d_1 \in \partial f_1$ and $d_2 \in \partial f_2$
- Composing with affine functions: $\partial f(Ax + b) = A^T \partial f(z)$ where $z = Ax + b$
- max: convex combinations of argmax gradients

$$\partial \max(f_1(x), f_2(x)) = \begin{cases} \nabla f_1(x) & \text{if } f_1(x) > f_2(x), \\ \nabla f_2(x) & \text{if } f_1(x) < f_2(x), \\ \nabla \theta f_1(x) + (1-\theta) f_2(x) & \text{if } f_1(x) = f_2(x), \end{cases}$$

$|x| = \max(x, -x)$

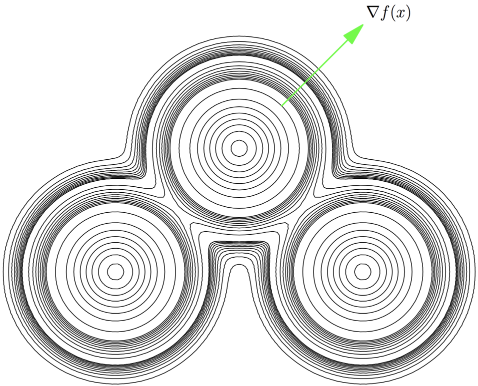
$\theta \nabla f_1(x)$

where $\theta \in [0, 1]$.

Subgradient Descent

Gradient orthogonal to level sets

We know that gradient points to the fastest ascent direction. What about subgradients?



Plot courtesy of Brett Bernstein.

Contour Lines and Subgradients

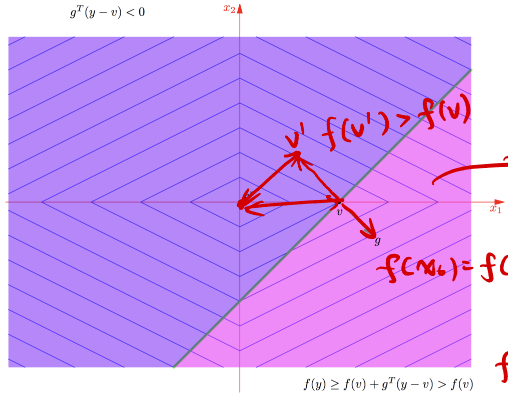
A hyperplane H **supports** a set S if H intersects S and all of S lies one one side of H.

Claim: If $f : \mathbb{R}^d \to \mathbb{R}$ has subgradient g at x_0, then the hyperplane H orthogonal to g at x_0 must **support** the level set $S = \left\{x \in \mathbb{R}^d \mid f(x) = f(x_0)\right\}$.

Proof:

- For any y, we have $f(y) \geqslant f(x_0) + g^T(y - x_0)$. (def of subgradient)
- If y is strictly on side of H that g points in,
  - then $g^T(y - x_0) > 0$.
  - So $f(y) > f(x_0)$.
  - So y is not in the level set S.
- $\therefore$ All elements of S must be on H or on the $-g$ side of H.

Subgradient of $f(x_1, x_2) = |x_1| + 2|x_2|$



$\rightarrow f(x) > f(v)$

$f(x_0) = f(v)$

$(x_0 - v) \cdot g > 0$

$f(x_0) \geq f(v) + \underbrace{g^T(x_0 - v)}_{>0}$

$f(x_0) > f(v)$

- Points on g side of H have larger f-values than $f(x_0)$. (from proof)
- But points on $-g$ side may **not** have smaller f-values.
- So $-g$ may **not** be a descent direction. (shown in figure)

Plot courtesy of Brett Bernstein.

Subgradient Descent

- Move along the negative subgradient:

$$x^{t+1} = x^t - \eta g \quad \text{where } g \in \partial f(x^t) \text{ and } \eta > 0$$

- This can increase the objective but gets us closer to the minimizer if f is convex and η is small enough:

$$\|x^{t+1} - x^*\| < \|x^t - x^*\|$$

- Subgradients don't necessarily converge to zero as we get closer to x^*, so we need decreasing step sizes, e.g. $O(1/t)$ or $O(1/\sqrt{t})$.

- Subgradient methods are slower than gradient descent, e.g. $O(1/\epsilon^2)$ vs $O(1/\epsilon)$ for convex functions.

Based on https://www.cs.ubc.ca/~schmidtm/Courses/5XX-S20/S4.pdf

Subgradient descent for SVM (HW3)

SVM objective function:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i w^T x_i\right) + \lambda \|w\|^2.$$

Pegasos: stochastic subgradient descent with step size $\eta_t = 1/(t\lambda)$

Input: $\lambda > 0$. Choose $w_1 = 0, t = 0$
While termination condition not met
  For $j = 1, \ldots, n$ (assumes data is randomly permuted)
    $t = t + 1$
    $\eta_t = 1/(t\lambda)$;
    If $y_j w_t^T x_j < 1$
      $w_{t+1} = (1 - \eta_t \lambda) w_t + \eta_t y_j x_j$ $w_{t+1} = w_t + y_j x_j$
    Else
      $w_{t+1} = (1 - \eta_t \lambda) w_t$

Summary

- Subgradient: generalize gradient for non-differentiable convex functions
- Subgradient "descent":
  - General method for non-smooth functions
  - Simple to implement
  - Slow to converge