Subgradient Descent

David S. Rosenberg

New York University

February 7, 2018
Motivation and Review: Support Vector Machines
The Classification Problem

- Output space $Y = \{-1, 1\}$  
  Action space $A = \mathbb{R}$
- **Real-valued prediction function** $f : \mathcal{X} \to \mathbb{R}$
- The value $f(x)$ is called the **score** for the input $x$.
- Intuitively, magnitude of the score represents the **confidence of our prediction**.
- Typical convention:
  
  $f(x) > 0$ $\implies$ Predict 1
  $f(x) < 0$ $\implies$ Predict $-1$

  (But we can choose other thresholds...)
The margin (or functional margin) for predicted score \( \hat{y} \) and true class \( y \in \{-1, 1\} \) is \( y\hat{y} \).

The margin often looks like \( yf(x) \), where \( f(x) \) is our score function.

The margin is a measure of how correct we are.

We want to maximize the margin.
[Margin-Based] Classification Losses

SVM/Hinge loss: \( \ell_{\text{Hinge}} = \max\{1 - m, 0\} = (1 - m)_+ \)

Not differentiable at \( m = 1 \). We have a "margin error" when \( m < 1 \).
Hypothesis space $\mathcal{F} = \{ f(x) = w^T x \mid w \in \mathbb{R}^d \}$.  

Loss $\ell(m) = \max(0, 1 - m)$  

$l_2$ regularization 

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \max(0, 1 - y_i w^T x_i) + \lambda \|w\|^2_2$$
SVM Optimization Problem (no intercept)

- SVM objective function:

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

- Not differentiable... but let’s think about gradient descent anyway.

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

\[
\ell'(m) = \begin{cases} 
0 & m > 1 \\
-1 & m < 1 \\
\text{undefined} & m = 1
\end{cases}
\]
“Gradient” of SVM Objective

- We need gradient with respect to parameter vector $w \in \mathbb{R}^d$:

$$
\nabla_w \ell (y_i w^T x_i) = \ell' (y_i w^T x_i) y_i x_i \text{ (chain rule)}
$$

$$
= \begin{cases} 
0 & y_i w^T x_i > 1 \\
-1 & y_i w^T x_i < 1 \\
\text{undefined} & y_i w^T x_i = 1
\end{cases} y_i x_i \text{ (expanded } m \text{ in } \ell'(m))
$$

$$
= \begin{cases} 
0 & y_i w^T x_i > 1 \\
-y_i x_i & y_i w^T x_i < 1 \\
\text{undefined} & y_i w^T x_i = 1
\end{cases}
$$
“Gradient” of SVM Objective

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

So

\[
\nabla_w J(w) = \nabla_w \left( \frac{1}{n} \sum_{i=1}^{n} \ell (y_i w^T x_i) + \lambda \|w\|^2 \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \nabla_w \ell (y_i w^T x_i) + 2\lambda w
\]

\[
= \begin{cases} 
\frac{1}{n} \sum_{i:y_i w^T x_i < 1} (-y_i x_i) + 2\lambda w & \text{all } y_i w^T x_i \neq 1 \\
\text{undefined} & \text{otherwise}
\end{cases}
\]
Gradient Descent on SVM Objective?

- The gradient of the SVM objective is

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

when $y_iw^T x_i \neq 1$ 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 $y_iw^T x_i = 1$?
- If we did, could we perturb the step size by $\varepsilon$ to miss such a point?
- Does it even make sense to check $y_iw^T x_i = 1$ with floating point numbers?
If we blindly apply gradient descent from a random starting point, it seems unlikely that we'll hit a point where the gradient is undefined.

Still, doesn't mean that gradient descent will work if objective not differentiable!

Theory of subgradients and subgradient descent will clear up any uncertainty.
Convexity and Sublevel Sets
Convex Sets

Definition

A set $C$ is **convex** if the line segment between any two points in $C$ lies in $C$. 

KPM Fig. 7.4
Convex and Concave Functions

Definition

A function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is convex if the line segment connecting any two points on the graph of \( f \) lies above the graph. \( f \) is concave if \(-f\) is convex.
Examples of Convex Functions on $\mathbb{R}$

Examples

- $x \mapsto ax + b$ is both convex and concave on $\mathbb{R}$ for all $a, b \in \mathbb{R}$.
- $x \mapsto |x|^p$ for $p \geq 1$ is convex on $\mathbb{R}$
- $x \mapsto e^{ax}$ is convex on $\mathbb{R}$ for all $a \in \mathbb{R}$
- Every norm on $\mathbb{R}^n$ is convex (e.g. $\|x\|_1$ and $\|x\|_2$)
- Max: $(x_1, \ldots, x_n) \mapsto \max\{x_1, \ldots, x_n\}$ is convex on $\mathbb{R}^n$
Simple Composition Rules

Examples

- If $g$ is convex, and $Ax + b$ is an affine mapping, then $g(Ax + b)$ is convex.
- If $g$ is convex then $\exp g(x)$ is convex.
- If $g$ is convex and nonnegative and $p \geq 1$ then $g(x)^p$ is convex.
- If $g$ is concave and positive then $\log g(x)$ is concave
- If $g$ is concave and positive then $1/g(x)$ is convex.
Main Reference for Convex Optimization

- Boyd and Vandenberghe (2004)
  - Very clearly written, but has a ton of detail for a first pass.
  - See the Extreme Abridgement of Boyd and Vandenberghe.
Convex Optimization Problem: Standard Form

minimize $f_0(x)$
subject to $f_i(x) \leq 0, \ i = 1, \ldots, m$

where $f_0, \ldots, f_m$ are convex functions.

Question: Is the $\leq$ in the constraint just a convention? Could we also have used $\geq$ or $=$?
Level Sets and Sublevel Sets

Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a function. Then we have the following definitions:

**Definition**

A **level set** or **contour line** for the value \( c \) is the set of points \( x \in \mathbb{R}^d \) for which \( f(x) = c \).

**Definition**

A **sublevel** set for the value \( c \) is the set of points \( x \in \mathbb{R}^d \) for which \( f(x) \leq c \).

**Theorem**

*If \( f : \mathbb{R}^d \to \mathbb{R} \) is convex, then the sublevel sets are convex.*

(Proof straight from definitions.)
Convex Function

Plot courtesy of Brett Bernstein.
Is the sublevel set \( \{ x \mid f(x) \leq 1 \} \) convex?
Nonconvex Function

Plot courtesy of Brett Bernstein.
Is the sublevel set \( \{ x \mid f(x) \leq 1 \} \) convex?

\[ f(x) \leq -1 \]
Fact: Intersection of Convex Sets is Convex

Plot courtesy of Brett Bernstein.

David S. Rosenberg  (New York University)
Level and Superlevel Sets

Level sets and superlevel sets of convex functions are **not** generally convex.

Plot courtesy of Brett Bernstein.
Convex Optimization Problem: Standard Form

minimize $f_0(x)$
subject to $f_i(x) \leq 0, \ i = 1, \ldots, m$

where $f_0, \ldots, f_m$ are convex functions.

- What can we say about each constraint set $\{x \mid f_i(x) \leq 0\}$? (convex)
- What can we say about the feasible set $\{x \mid f_i(x) \leq 0, \ i = 1, \ldots, m\}$? (convex)
Convex Optimization Problem: Implicit Form

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in C
\end{align*}
\]

where \( f \) is a convex function and \( C \) is a convex set.

An alternative “generic” convex optimization problem.
Convex and Differentiable Functions
First-Order Approximation

- Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable.
- Predict $f(y)$ given $f(x)$ and $\nabla f(x)$?
- Linear (i.e. “first order”) approximation:

$$f(y) \approx f(x) + \nabla f(x)^T (y - x)$$
First-Order Condition for Convex, Differentiable Function

- Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable.
- Then for any $x, y \in \mathbb{R}^d$ 
  \[ f(y) \geq f(x) + \nabla f(x)^T (y - x) \]
- The linear approximation to $f$ at $x$ is a **global underestimator** of $f$:

---

Figure from Boyd & Vandenberghe Fig. 3.2; Proof in Section 3.1.3
First-Order Condition for Convex, Differentiable Function

- Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable
- Then for any $x, y \in \mathbb{R}^d$
  \[ f(y) \geq f(x) + \nabla f(x)^T (y - x) \]

**Corollary**

*If $\nabla f(x) = 0$ then $x$ is a global minimizer of $f$.\*

For convex functions, **local information gives global information.**
Subgradients
Subgradients

Definition

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

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

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 $f(x)$. 
Subdifferential

Definitions

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

Basic Facts

- $f$ is convex and differentiable $\implies \partial f(x) = \{\nabla f(x)\}$.
- Any point $x$, there can be 0, 1, or infinitely many subgradients.
- $\partial f(x) = \emptyset \implies f$ is not convex.
Globla Optimality Condition

Definition
A vector \( g \in \mathbb{R}^d \) is a subgradient of \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) at \( x \) if for all \( z \),

\[
f(z) \geq f(x) + g^T(z - x).
\]

Corollary
If \( 0 \in \partial f(x) \), then \( x \) is a global minimizer of \( f \).
Subdifferential of Absolute Value

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

- Plot on right shows $\{(x, g) \mid x \in \mathbb{R}, g \in \partial f(x)\}$
\[ f(x_1, x_2) = |x_1| + 2|x_2| \]
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 $g_2 \in [-2, 2]$. 
Underestimating Hyperplane to $f(x_1, x_2) = |x_1| + 2|x_2|$
Contour plot of $f(x_1, x_2) = |x_1| + 2|x_2|$, with set of subgradients at $(3, 0)$. 

\[ \partial f(3, 0) = \{(1, b)^T \mid b \in [-2, 2]\} \]
Contour Lines and Gradients

- For function $f : \mathbb{R}^d \to \mathbb{R}$,
  - **graph** of function lives in $\mathbb{R}^{d+1}$,
  - **gradient** and **subgradient** of $f$ live in $\mathbb{R}^d$, and
  - **contours**, **level sets**, and **sublevel sets** are in $\mathbb{R}^d$.

- $f : \mathbb{R}^d \to \mathbb{R}$ continuously differentiable, $\nabla f(x_0) \neq 0$, then $\nabla f(x_0)$ normal to level set

  $$S = \{ x \in \mathbb{R}^d \mid f(x) = f(x_0) \}.$$

- Proof sketch in notes.
Gradient orthogonal to sublevel sets

\[ \nabla f(x) \]