# Subgradient Descent 

David S. Rosenberg<br>Bloomberg ML EDU

October 18, 2017

Motivation and Review: Support Vector Machines

## The Classification Problem

- Output space $y=\{-1,1\} \quad$ Action space $\mathcal{A}=\mathbf{R}$
- Real-valued prediction function $f: X \rightarrow \mathbf{R}$
- 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:

$$
\begin{aligned}
& f(x)>0 \Longrightarrow \text { Predict 1 } \\
& f(x)<0 \Longrightarrow \text { Predict -1 }
\end{aligned}
$$

(But we can choose other thresholds...)

## The Margin

- 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 $y f(x)$, 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 $m<1$.

## [Soft Margin] Linear Support Vector Machine (No Intercept)

- Hypothesis space $\mathcal{F}=\left\{f(x)=w^{T} x \mid w \in \mathbf{R}^{d}\right\}$.
- Loss $\ell(m)=\max (1, m)$
- $\ell_{2}$ regularization

$$
\min _{w \in \mathbf{R}^{d}} \sum_{i=1}^{n} \max \left(0,1-y_{i} w^{T} x_{i}\right)+\lambda\|w\|_{2}^{2}
$$

## SVM Optimization Problem (no intercept)

- SVM objective function:

$$
J(w)=\frac{1}{n} \sum_{i=1}^{n} \max \left(0,1-y_{i}\left[w^{\top} x_{i}\right]\right)+\lambda\|w\|^{2} .
$$

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

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

## "Gradient" of SVM Objective

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

$$
\begin{aligned}
\nabla_{w} \ell\left(y_{i} w^{\top} x_{i}\right) & =\ell^{\prime}\left(y_{i} w^{\top} x_{i}\right) y_{i} x_{i} \text { (chain rule) } \\
& \left.=\left(\begin{array}{ll}
0 & y_{i} w^{\top} x_{i}>1 \\
-1 & y_{i} w^{\top} x_{i}<1 \\
\text { undefined } & y_{i} w^{\top} x_{i}=1
\end{array}\right) y_{i} x_{i} \text { (expanded } m \text { in } \ell^{\prime}(m)\right) \\
& = \begin{cases}0 & y_{i} w^{\top} x_{i}>1 \\
-y_{i} x_{i} & y_{i} w^{\top} x_{i}<1 \\
\text { undefined } & y_{i} w^{\top} x_{i}=1\end{cases}
\end{aligned}
$$

## "Gradient" of SVM Objective

$$
\nabla_{w} \ell\left(y_{i} w^{T} x_{i}\right)= \begin{cases}0 & y_{i} w^{T} x_{i}>1 \\ -y_{i} x_{i} & y_{i} w^{\top} x_{i}<1 \\ \text { undefined } & y_{i} w^{\top} x_{i}=1\end{cases}
$$

So

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

## Gradient Descent on SVM Objective?

- The gradient of the SVM objective is

$$
\nabla_{w} J(w)=\frac{1}{n} \sum_{i: y_{i} w^{T} x_{i}<1}\left(-y_{i} x_{i}\right)+2 \lambda w
$$

when $y_{i} w^{T} x_{i} \neq 1$ for all $i$, and otherwise is undefined.
Suppose we tried gradient descent on $J(w)$ :

- If we start with a random $w$, will we ever hit $y_{i} w^{\top} x_{i}=1$ ?
- If we did, could we perturb the step size by $\varepsilon$ to miss such a point?
- Does it even make sense to check $y_{i} w^{\top} x_{i}=1$ with floating point numbers?


## Gradient Descent on SVM Objective?

- If we blindly apply gradient descent from a random starting point
- 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 $C$.


## Convex and Concave Functions

## Definition

A function $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ 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.


KPM Fig. 7.5

## Convex Optimization Problem: Standard Form

Convex Optimization Problem: Standard Form

$$
\begin{array}{ll}
\operatorname{minimize} & f_{0}(x) \\
\text { subject to } & f_{i}(x) \leqslant 0, \quad i=1, \ldots, m
\end{array}
$$

where $f_{0}, \ldots, f_{m}$ are convex functions.
Question: Is the $\leqslant$ in the constraint just a convention? Could we also have used $\geqslant$ or $=$ ?

## Level Sets and Sublevel Sets

Let $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ 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 $x \in \mathbf{R}^{d}$ for which $f(x)=c$.

Definition
A sublevel set for the value $c$ is the set of points $x \in \mathbf{R}^{d}$ for which $f(x) \leqslant c$.

Theorem
If $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ is convex, then the sublevel sets are convex.
(Proof straight from definitions.)

## Convex Function



Plot courtesy of Brett Bernstein.

## Contour Plot Convex Function: Sublevel Set



Is the sublevel set $\{x \mid f(x) \leqslant 1\}$ convex?

## Nonconvex Function



Plot courtesy of Brett Bernstein.

## Contour Plot Nonconvex Function: Sublevel Set



Is the sublevel set $\{x \mid f(x) \leqslant 1\}$ convex?

## Fact: Intersection of Convex Sets is Convex



Plot courtesy of Brett Bernstein.

## Level and Superlevel Sets



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

## Convex Optimization Problem: Standard Form

Convex Optimization Problem: Standard Form

| minimize | $f_{0}(x)$ |
| :---: | :--- |
| subject to | $f_{i}(x) \leqslant 0, \quad i=1, \ldots, m$ |

where $f_{0}, \ldots, f_{m}$ are convex functions.

- What can we say about each constraint set $\left\{x \mid f_{i}(x) \leqslant 0\right\}$ ? (convex)
- What can we say about the feasible set $\left\{x \mid f_{i}(x) \leqslant 0, i=1, \ldots, m\right\}$ ? (convex)


## Convex Optimization Problem: Implicit Form

Convex Optimization Problem: Implicit Form

| $\operatorname{minimize}$ | $f(x)$ |
| :--- | :--- |
| subject to | $x \in C$ |

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

$$
f(y) \approx f(x)+\nabla f(x)^{T}(y-x)
$$



Boyd \& Vandenberghe Fig. 3.2

## First-Order Condition for Convex, Differentiable Function

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

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

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


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

## First-Order Condition for Convex, Differentiable Function

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

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

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 $g \in \mathbf{R}^{d}$ is a subgradient of $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ 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 $x \mapsto f\left(x_{0}\right)+g^{T}\left(x-x_{0}\right)$ 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 $\Longrightarrow \partial f(x)=\{\nabla f(x)\}$.
- Any point $x$, there can be 0,1 , or infinitely many subgradients.
- $\partial f(x)=\emptyset \Longrightarrow f$ is not convex.


## Globla Optimality Condition

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

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

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 $\{(x, g) \mid x \in \mathbf{R}, g \in \partial f(x)\}$


## $f\left(x_{1}, x_{2}\right)=\left|x_{1}\right|+2\left|x_{2}\right|$



## Subgradients of $f\left(x_{1}, x_{2}\right)=\left|x_{1}\right|+2\left|x_{2}\right|$

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


## Underestimating Hyperplane to $f\left(x_{1}, x_{2}\right)=\left|x_{1}\right|+2\left|x_{2}\right|$



Plot courtesy of Brett Bernstein.

## Subdifferential on Contour Plot

$$
\partial f(3,0)=\left\{(1, b)^{T} \mid b \in[-2,2]\right\}
$$



Contour plot of $f\left(x_{1}, x_{2}\right)=\left|x_{1}\right|+2\left|x_{2}\right|$, with set of subgradients at $(3,0)$.

## Contour Lines and Gradients

- For function $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$,
- graph of function lives in $\mathbf{R}^{d+1}$,
- gradient and subgradient of $f$ live in $\mathbf{R}^{d}$, and
- contours, level sets, and sublevel sets are in $\mathbf{R}^{d}$.
- $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ continuously differentiable, $\nabla f\left(x_{0}\right) \neq 0$, then $\nabla f\left(x_{0}\right)$ normal to level set

$$
S=\left\{x \in \mathbf{R}^{d} \mid f(x)=f\left(x_{0}\right)\right\} .
$$

- Proof sketch in notes.


## Gradient orthogonal to sublevel sets



Plot courtesy of Brett Bernstein.

## Contour Lines and Subgradients

Let $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ have a subgradient $g$ at $x_{0}$.

- Hyperplane $H$ orthogonal to $g$ at $x_{0}$ must support the level set $S=\left\{x \in \mathbf{R}^{d} \mid f(x)=f\left(x_{0}\right)\right\}$.
- i.e $H$ contains $x_{0}$ and all of $S$ lies one one side of $H$.

Proof:

- For any $y$, we have $f(y) \geqslant f\left(x_{0}\right)+g^{T}\left(y-x_{0}\right)$. (def of subgradient)
- If $y$ is strictly on side of $H$ that $g$ points in,
- then $g^{T}\left(y-x_{0}\right)>0$.
- So $f(y)>f\left(x_{0}\right)$.
- 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 $f\left(x_{1}, x_{2}\right)=\left|x_{1}\right|+2\left|x_{2}\right|$



Plot courtesy of Brett Bernstein.

## Subgradient of $f\left(x_{1}, x_{2}\right)=\left|x_{1}\right|+2\left|x_{2}\right|$



- Points on $g$ side of $H$ have larger $f$-values than $f\left(x_{0}\right)$. (from proof)
- But points on $-g$ side may not have smaller $f$-values.
- So -g may not be a descent direction. (shown in figure)


## Subgradient Descent

## Subgradient Descent

- Suppose $f$ is convex, and we start optimizing at $x_{0}$.
- Repeat
- Step in a negative subgradient direction:

$$
x=x_{0}-t g,
$$

where $t>0$ is the step size and $g \in \partial f\left(x_{0}\right)$.
$-g$ not a descent direction - can this work?

## Subgradient Gets Us Closer To Minimizer

## Theorem

Suppose $f$ is convex.

- Let $x=x_{0}-t g$, for $g \in \partial f\left(x_{0}\right)$.
- Let $z$ be any point for which $f(z)<f\left(x_{0}\right)$.
- Then for small enough $t>0$,

$$
\|x-z\|_{2}<\left\|x_{0}-z\right\|_{2} .
$$

- Apply this with $z=x^{*} \in \arg \min _{x} f(x)$.
$\Longrightarrow$ Negative subgradient step gets us closer to minimizer.


## Subgradient Gets Us Closer To Minimizer (Proof)

- Let $x=x_{0}-t g$, for $g \in \partial f\left(x_{0}\right)$ and $t>0$.
- Let $z$ be any point for which $f(z)<f\left(x_{0}\right)$.
- Then

$$
\begin{aligned}
\|x-z\|_{2}^{2} & =\left\|x_{0}-t g-z\right\|_{2}^{2} \\
& =\left\|x_{0}-z\right\|_{2}^{2}-2 \operatorname{tg}^{T}\left(x_{0}-z\right)+t^{2}\|g\|_{2}^{2} \\
& \leqslant\left\|x_{0}-z\right\|_{2}^{2}-2 t\left[f\left(x_{0}\right)-f(z)\right]+t^{2}\|g\|_{2}^{2}
\end{aligned}
$$

- Consider $-2 t\left[f\left(x_{0}\right)-f(z)\right]+t^{2}\|g\|_{2}^{2}$.
- It's a convex quadratic (facing upwards).
- Has zeros at $t=0$ and $t=2\left(f\left(x_{0}\right)-f(z)\right) /\|g\|_{2}^{2}>0$.
- Therefore, it's negative for any

$$
t \in\left(0, \frac{2\left(f\left(x_{0}\right)-f(z)\right)}{\|g\|_{2}^{2}}\right) .
$$

## Based on Boyd EE364b: Subgradients Slides

## Convergence Theorem for Fixed Step Size

Assume $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ is convex and

- $f$ is Lipschitz continuous with constant $G>0$ :

$$
|f(x)-f(y)| \leqslant G\|x-y\| \text { for all } x, y
$$

## Theorem

For fixed step size $t$, subgradient method satisfies:

$$
\lim _{k \rightarrow \infty} f\left(x_{\text {best }}^{(k)}\right) \leqslant f\left(x^{*}\right)+G^{2} t / 2
$$

## Convergence Theorems for Decreasing Step Sizes

Assume $f: \mathbf{R}^{d} \rightarrow \mathbf{R}$ is convex and

- $f$ is Lipschitz continuous with constant $G>0$ :

$$
|f(x)-f(y)| \leqslant G\|x-y\| \text { for all } x, y
$$

Theorem
For step size respecting Robbins-Monro conditions,

$$
\lim _{k \rightarrow \infty} f\left(x_{\text {best }}^{(k)}\right)=f\left(x^{*}\right)
$$

