Subgradient Descent

David S. Rosenberg

New York University

February 7, 2018

Contents

Motivation and Review: Support Vector Machines

The Classification Problem

- Output space $\mathcal{Y} = \{-1, 1\}$ Action space $\mathcal{A} = \mathbf{R}$
- **Real-valued prediction function $f : \mathcal{X} \to \mathbf{R}$**

- 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:

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

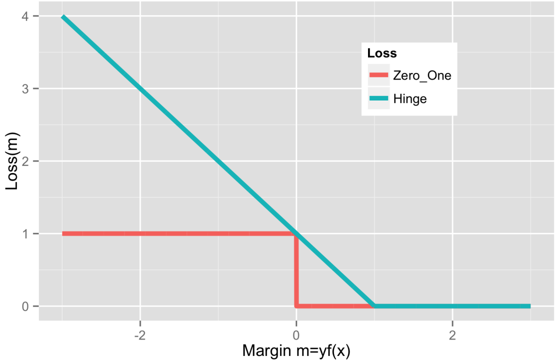
(But we can choose other thresholds...)

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 $yf(x)$, 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 $m < 1$.

[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 $\ell(m) = \max(0, 1 - m)$
- ℓ_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:

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

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

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

"Gradient" of SVM Objective

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

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

"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?

Gradient Descent on SVM Objective?

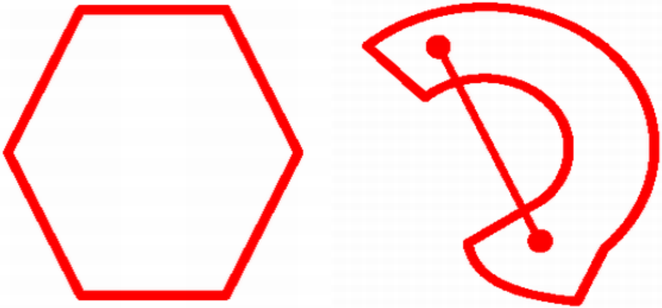
- If we blindly apply gradient descent from a random starting point
  - seems unlikely that we'll hit a point where the gradient is undefined.
- Still, doesn't mean that gradient descent will work if objective not differentiable!
- Theory of subgradients and subgradient descent will clear up any uncertainty.

Convexity and Sublevel Sets

Convex Sets

Definition

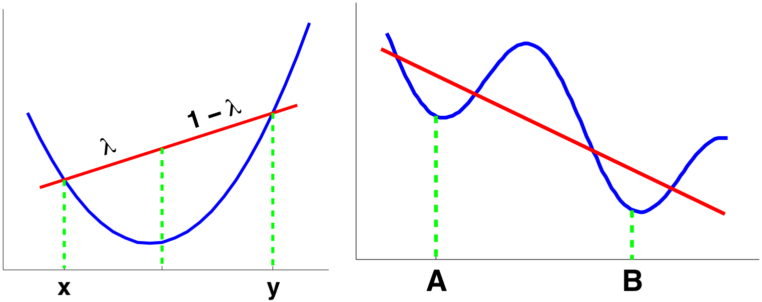
A set C is **convex** if the line segment between any two points in C lies in C.



KPM Fig. 7.4

Definition

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

Examples of Convex Functions on $\mathbf{R}$

Examples

- $x \mapsto ax + b$ is both convex and concave on $\mathbf{R}$ for all $a, b \in \mathbf{R}$.
- $x \mapsto |x|^p$ for $p \geqslant 1$ is convex on $\mathbf{R}$
- $x \mapsto e^{ax}$ is convex on $\mathbf{R}$ for all $a \in \mathbf{R}$
- Every norm on $\mathbf{R}^n$ 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 $\mathbf{R}^n$

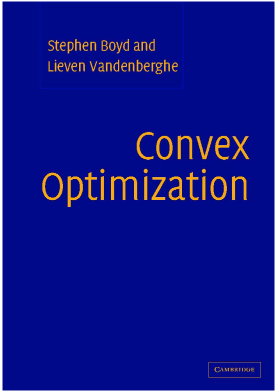
Simple Composition Rules

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 $p \geqslant 1$ 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

$$
\begin{array}{ll}
\text{minimize} & f_0(x) \\
\text{subject to} & f_i(x) \leqslant 0, \ \ i = 1, \ldots, m
\end{array}
$$

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 $f : \mathbf{R}^d \to \mathbf{R}$ 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

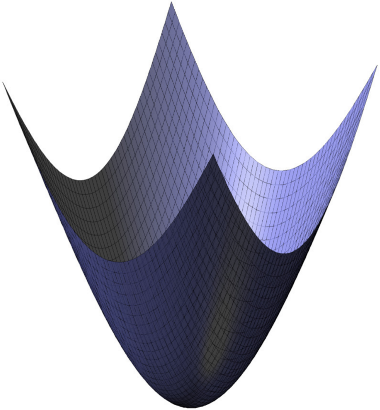
A **sublevel** set for the value c is the set of points $x \in \mathbf{R}^d$ for which $f(x) \leqslant c$.

Theorem

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

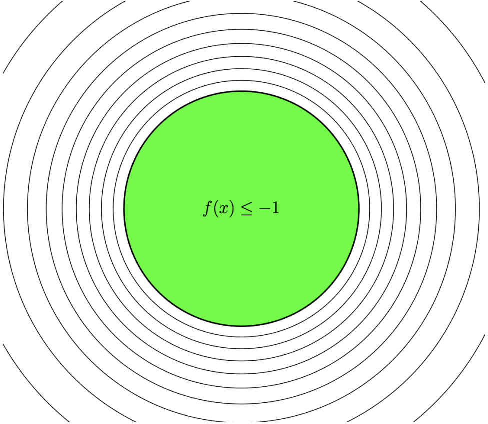
(Proof straight from definitions.)

Convex Function



Plot courtesy of Brett Bernstein.

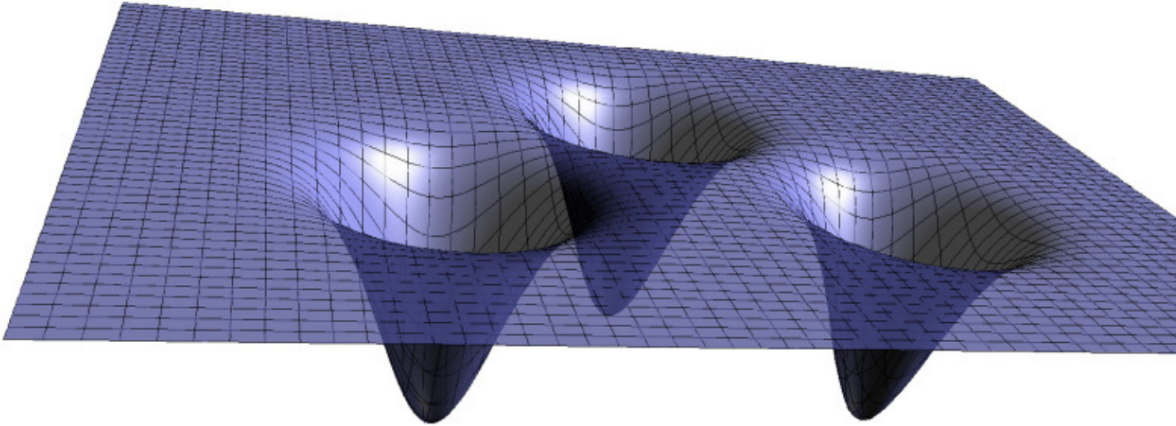
Contour Plot Convex Function: Sublevel Set



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

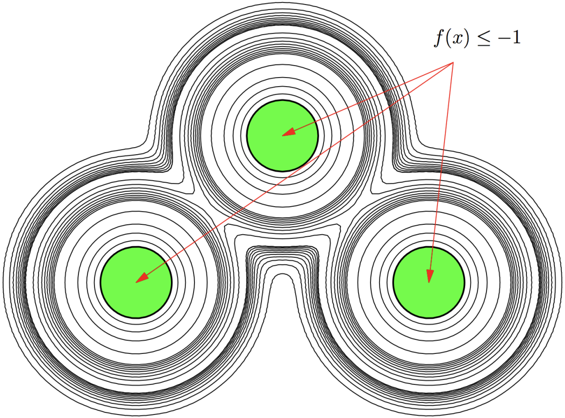
Plot courtesy of Brett Bernstein.

Nonconvex Function



Plot courtesy of Brett Bernstein.

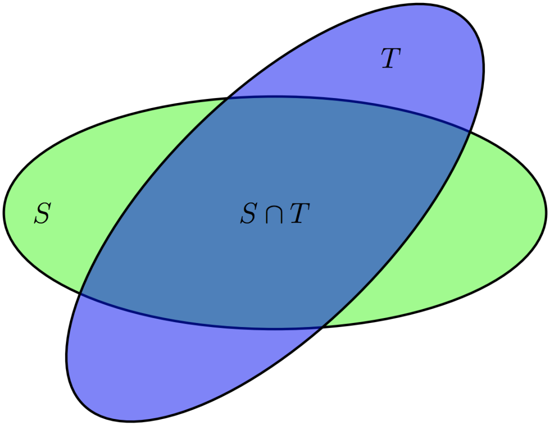
Contour Plot Nonconvex Function: Sublevel Set



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

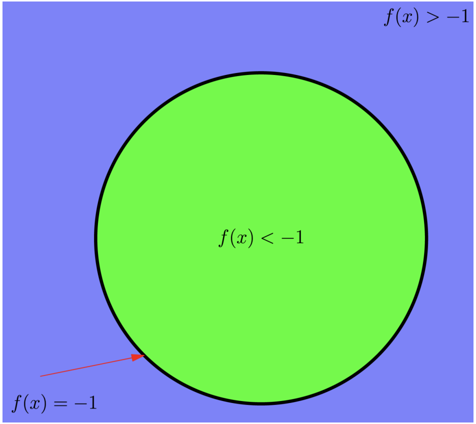
Plot courtesy of Brett Bernstein.

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.

Plot courtesy of Brett Bernstein.

Convex Optimization Problem: Standard Form

Convex Optimization Problem: Standard Form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leqslant 0, \;\; i = 1, \ldots, m \end{array}$$

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

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

Convex Optimization Problem: Implicit Form

Convex Optimization Problem: Implicit Form

$$
\begin{array}{ll}
\text{minimize} & f(x) \\
\text{subject to} & x \in C
\end{array}
$$

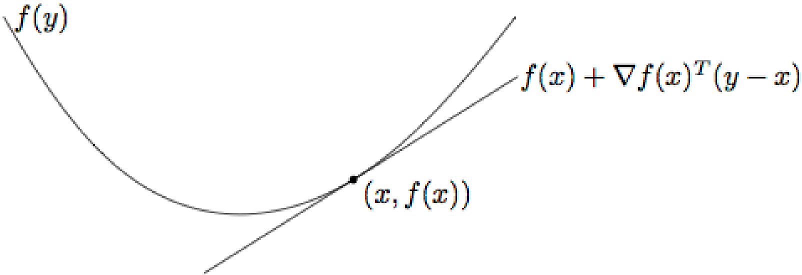
where f is a convex function and C is a convex set.
An alternative "generic" convex optimization problem.

Convex and Differentiable Functions

First-Order Approximation

- Suppose $f : \mathbf{R}^d \to \mathbf{R}$ 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 $f : \mathbf{R}^d \to \mathbf{R}$ 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:

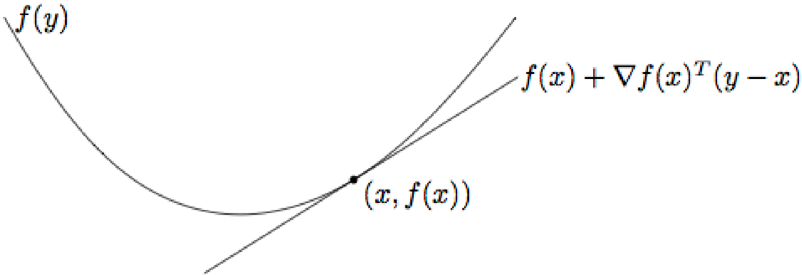


Figure from Boyd & Vandenberghe Fig. 3.2; Proof in Section 3.1.3

First-Order Condition for Convex, Differentiable Function

- Suppose $f : \mathbf{R}^d \to \mathbf{R}$ 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.**

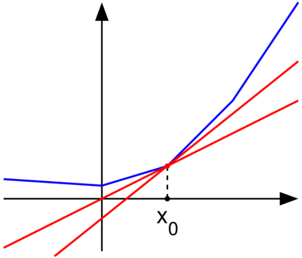
Subgradients

Subgradients

Definition

A vector $g \in \mathbf{R}^d$ is a **subgradient** of $f : \mathbf{R}^d \to \mathbf{R}$ 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 $f(x)$.

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

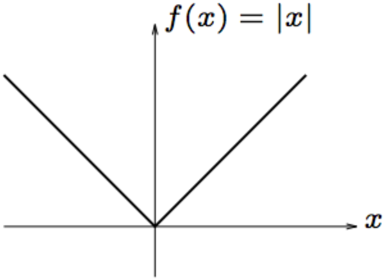
$$f(z) \geqslant f(x) + g^T(z - x).$$

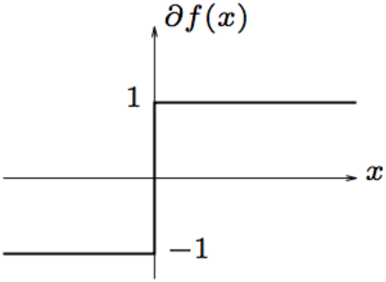
Corollary

If $0 \in \partial f(x)$, then x is a ***global minimizer*** *of f.*

Subdifferential of Absolute Value

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

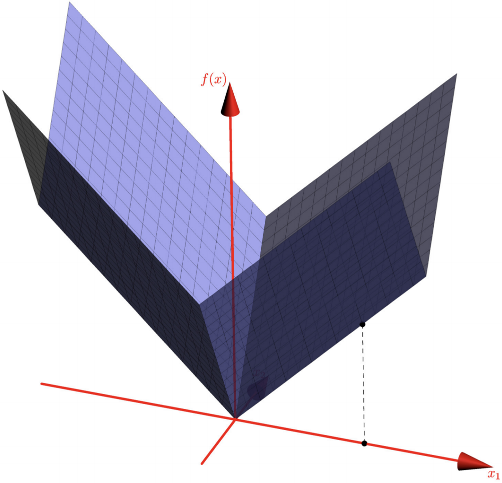




- Plot on right shows $\{(x, g) \mid x \in \mathbf{R},\ g \in \partial f(x)\}$

Boyd EE364b: Subgradients Slides

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

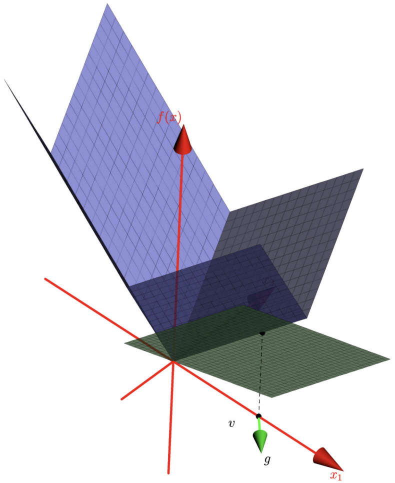


Plot courtesy of Brett Bernstein.

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 $g_2 \in [-2, 2]$.

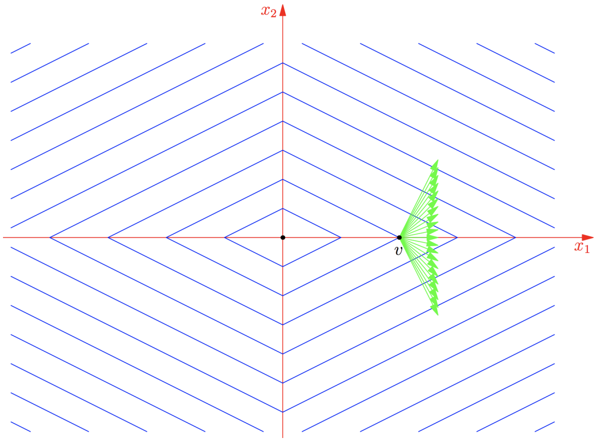
Underestimating Hyperplane to $f(x_1, x_2) = |x_1| + 2|x_2|$



Plot courtesy of Brett Bernstein.

Subdifferential on Contour Plot

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



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

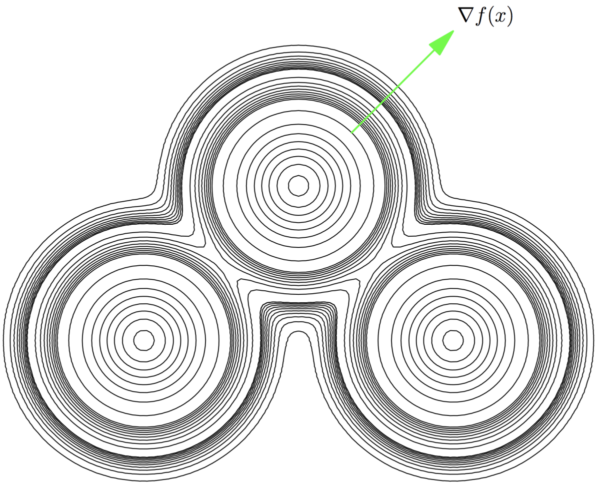
Contour Lines and Gradients

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

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

- Proof sketch in notes.

Gradient orthogonal to sublevel sets



Plot courtesy of Brett Bernstein.