Machine Learning – Brett Bernstein

Recitation 4: Subgradients

Intro Question

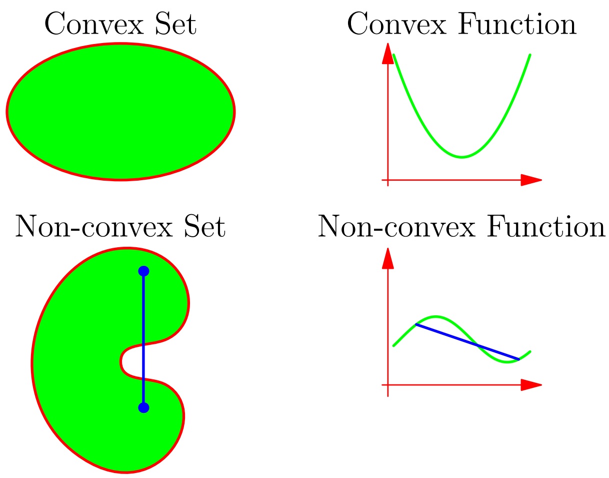
1. When stating a convex optimization problem in standard form we write

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0 \quad \text{for all } i = 1, \ldots, n. \end{array}$$

   where $f_0, f_1, \ldots, f_n$ are convex. Why don't we use $\ge$ or $=$ instead of $\le$?

More on Convexity and Review of Duality

Recall that a set $S \subseteq \mathbb{R}^d$ is convex if for any $x, y \in S$ and $\theta \in (0,1)$ we have $(1-\theta)x + \theta y \in S$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^d$ and $\theta \in (0,1)$ we have $f((1-\theta)x + \theta y) \le (1-\theta)f(x) + \theta f(y)$.

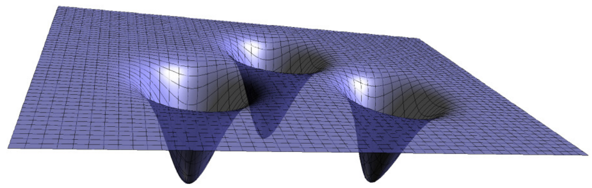


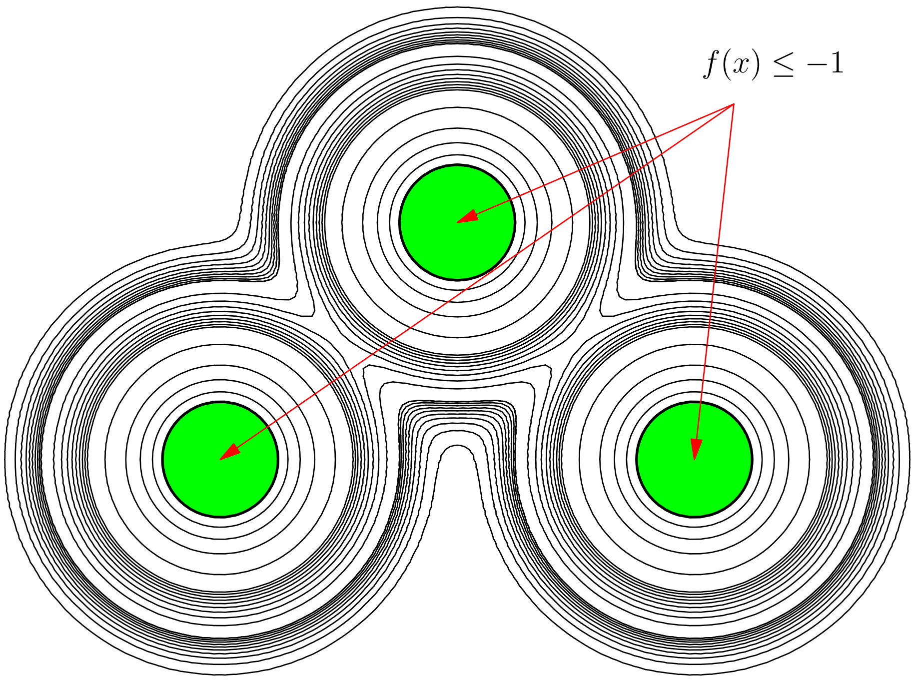
For a function $f : \mathbb{R}^d \to \mathbb{R}$, a level set (or contour line) corresponding to the value c is given by the set of all points $x \in \mathbb{R}^d$ where $f(x) = c$:

$$f^{-1}\{c\} = \{x \in \mathbb{R}^d \mid f(x) = c\}.$$

Analogously, the sublevel set for the value c is the set of all points $x \in \mathbb{R}^d$ where $f(x) \le c$:

$$f^{-1}(-\infty, c] = \{x \in \mathbb{R}^d \mid f(x) \le c\}.$$





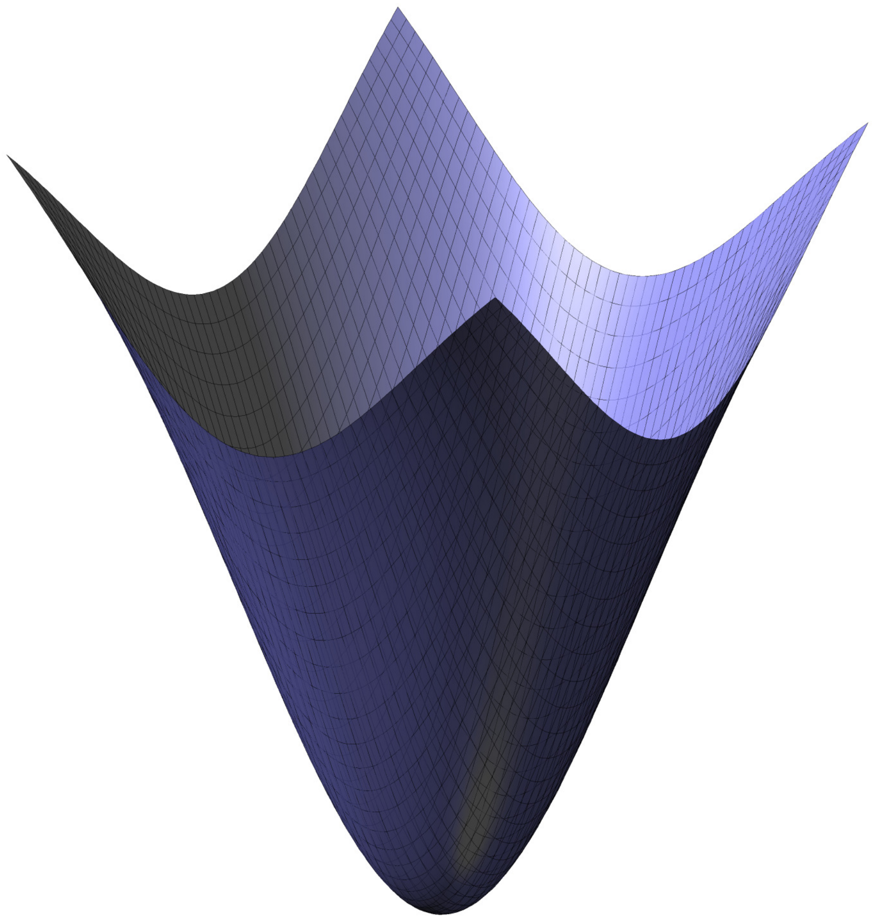
Above is a non-convex function, the contour plot, and the sublevel set where $f(x) \leq -1$. When f is convex, we can say something nice about these sets.

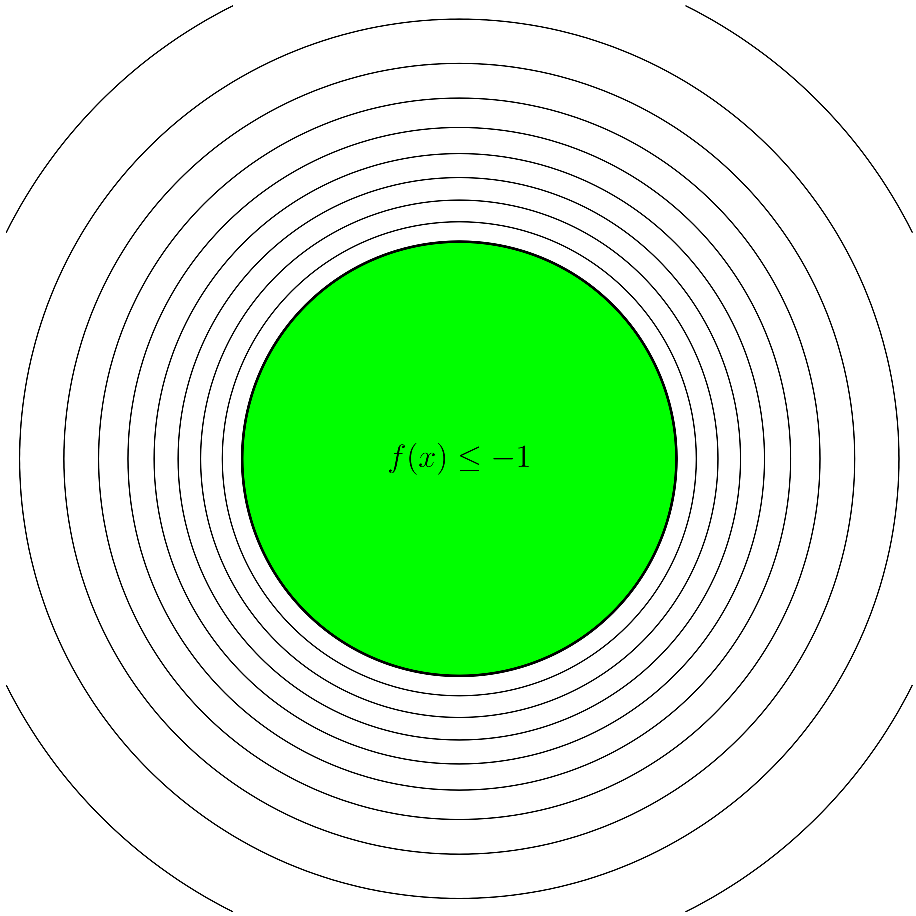
Theorem 1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex then the sublevel sets are convex.*

Proof. Fix a sublevel set $S = \{x \in \mathbb{R}^d \mid f(x) \leq c\}$ for some fixed $c \in \mathbb{R}$. If $x, y \in S$ and $\theta \in (0, 1)$ then we have

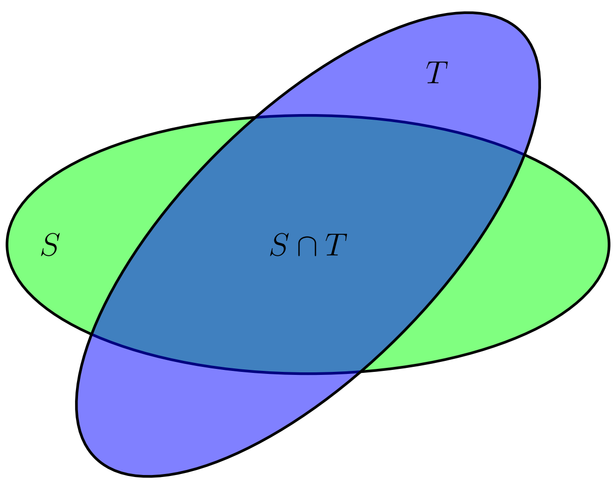
$$f((1-\theta)x + \theta y) \leq (1-\theta)f(x) + \theta f(y) \leq (1-\theta)c + \theta c = c.$$

□

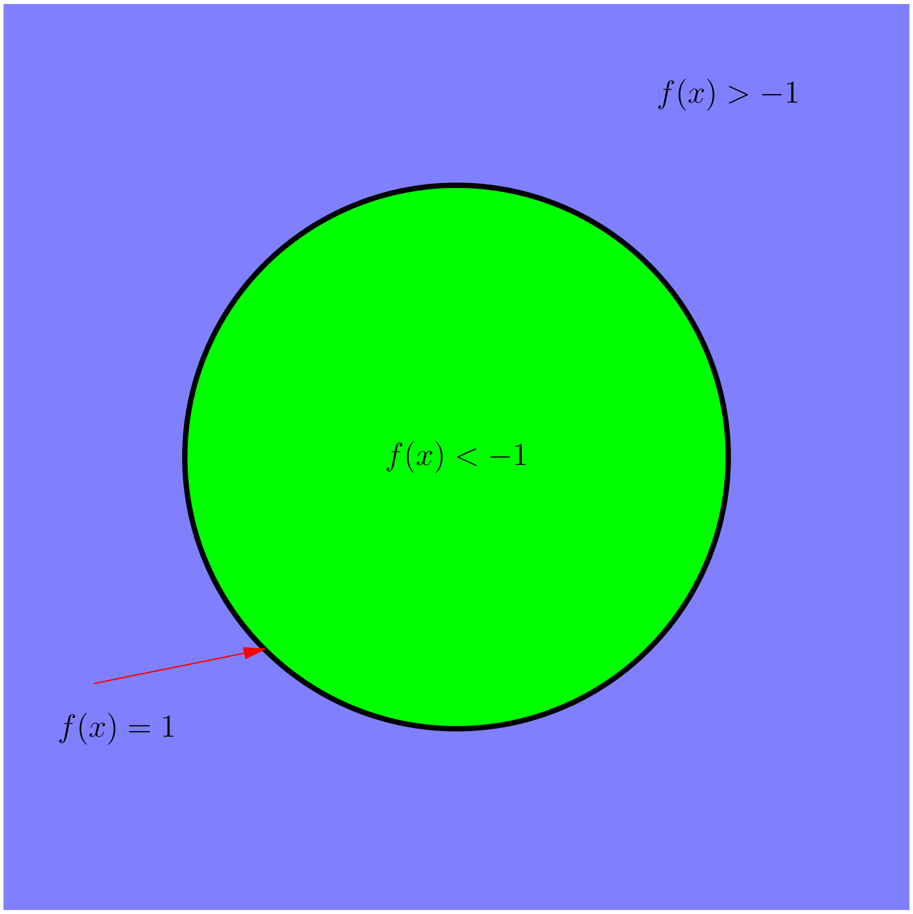




In the concept check questions we will show that the intersection of convex sets is convex.



This proves that having a bunch of conditions of the form $f_i(x) \leq 0$ where the f_i are convex gives us a convex feasible set. While the sublevel sets are convex, a convex function need not have convex level sets. Furthermore, sets of the form $\{x \in \mathbb{R}^d \mid f(x) \geq c\}$ also need not be convex (called superlevel sets).

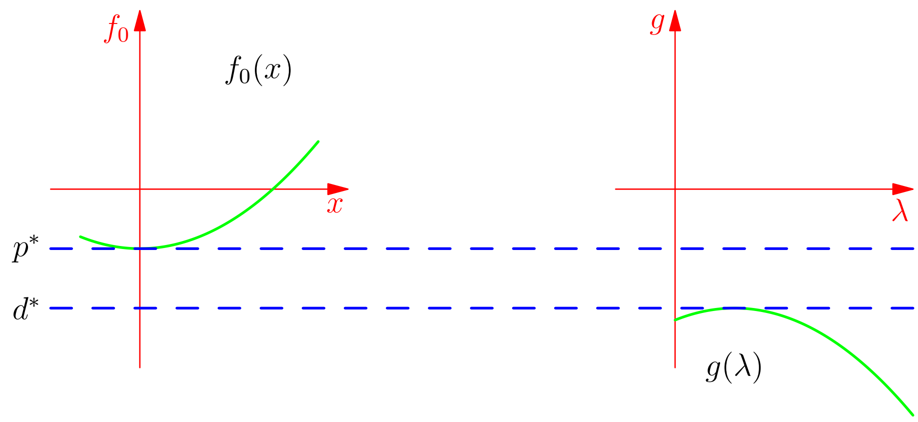


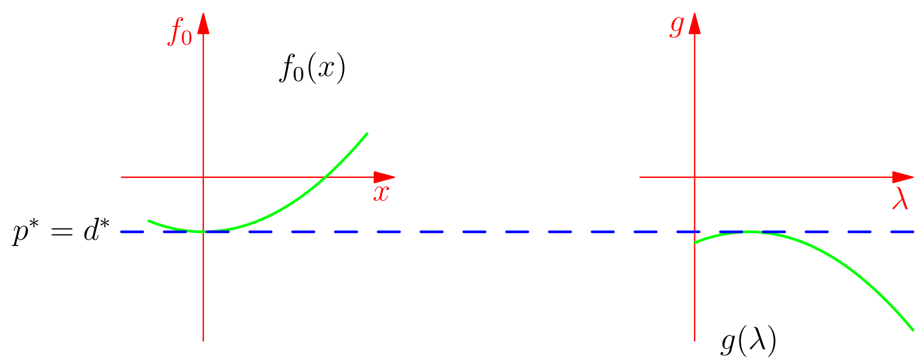
This brings us to the question, why do we care about convexity? Here are some reasons.

1. If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then local minima are global minima.
2. Given a point $x \in \mathbb{R}^d$ and a closed convex set S, there is a unique point of S that is closest to x (called the projection of x onto S).
3. A pair of disjoint convex sets can be separated by a hyperplane (used to prove Slater's condition for strong duality).

We also discussed duality as seen below. Lagrange duality let's us change our optimization problem into a new problem with potentially simpler constraints. Moreover, the Lagrange dual optimal value d^* will always be less than the primal optimal value p^* (called weak duality). If we satisfy certain conditions (Slater) we get strong duality ($p^* = d^*$). Using the strong duality relationship we can derive interesting relations between the primal and dual solutions (e.g., complementary slackness).

$$\boxed{\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0 \end{array}} \longrightarrow \boxed{\text{Lagrange Dual}} \longrightarrow \boxed{\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \geq 0 \end{array}}$$





Gradients and Subgradients

Definitions and Basic Properties

Recall that for differentiable $f : \mathbb{R}^d \to \mathbb{R}$ we can write the linear approximation

$$f(x + v) \approx f(x) + \nabla f(x)^T v,$$

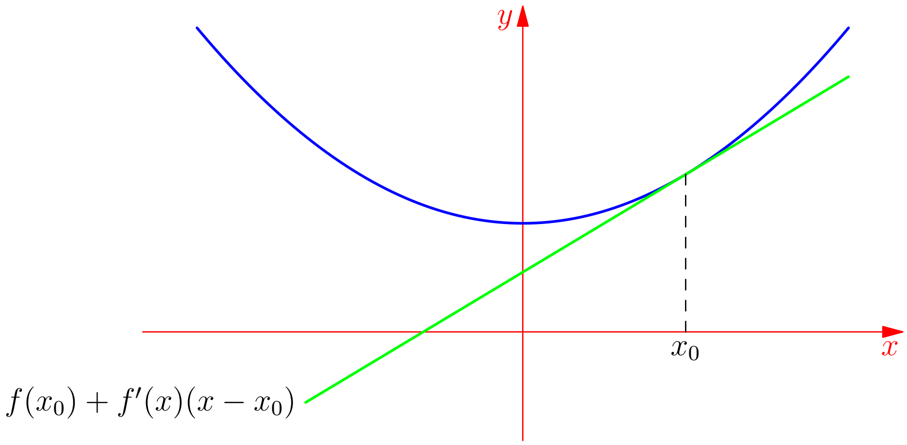
when v is small. We can use gradients to characterize convexity.

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then f is convex iff*

$$f(x + v) \geq f(x) + \nabla f(x)^T v$$

hold for all $x, v \in \mathbb{R}^d$.

In words, this says that the approximating tangent line (or hyperplane in higher dimensions) is a global underestimator (lies entirely below the function).



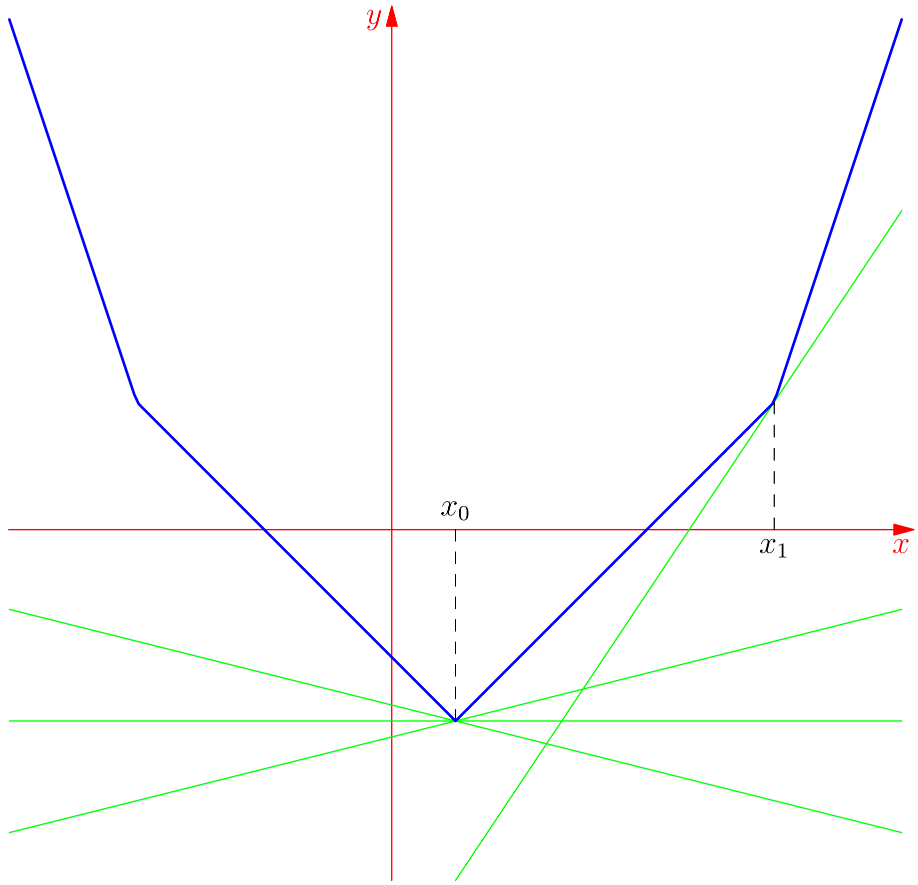
Even if f is not differentiable at x, we can still look for vectors satisfying a similar relationship.

Definition 3 (Subgradient, Subdifferential, Subdifferentiable)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$. We say that $g \in \mathbb{R}^d$ is a *subgradient* of f at $x \in \mathbb{R}^d$ if

$$f(x + v) \geq f(x) + g^T v$$

for all $v \in \mathbb{R}^d$. The *subdifferential* $\partial f(x)$ is the set of all subgradients of f at x. We say that f is *subdifferentiable* at x if $\partial f(x) \neq \emptyset$ (i.e., if there is at least one subgradient).

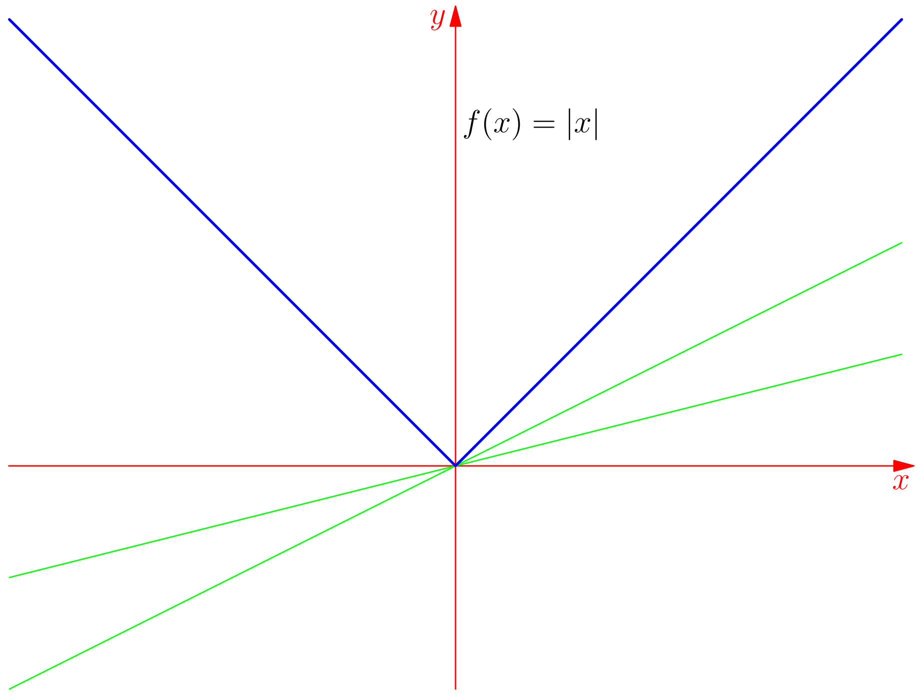
Below are subgradients drawn at x_0 and x_1.



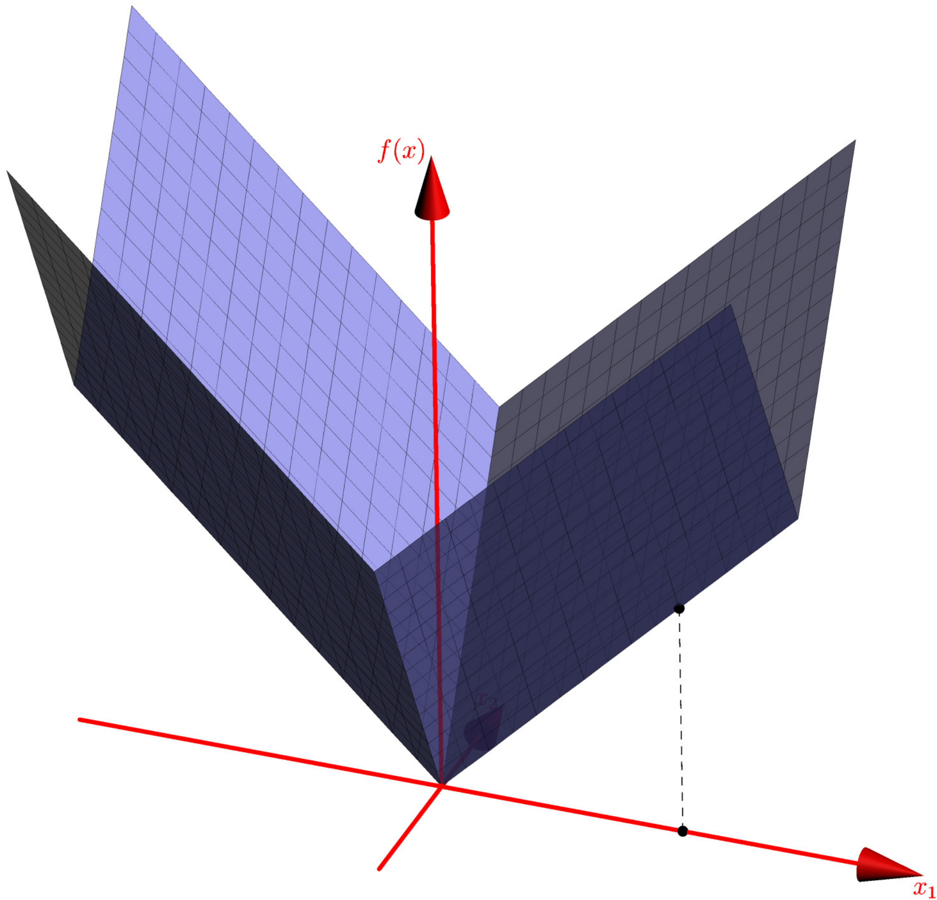
Facts about subgradients (proven in the concept check exercises).

1. If f is convex and differentiable at x then $\partial f(x) = \{\nabla f(x)\}$.
2. If f is convex then $\partial f(x) \neq \emptyset$ for all x.
3. The subdifferential $\partial f(x)$ is a convex set. Thus the subdifferential can contain 0, 1, or infinitely many elements.
4. If the zero vector is a subgradient of f at x, then x is a global minimum.
5. If g is a subgradient of f at x, then $(g, -1)$ is orthogonal to the underestimating hyperplane $\{(x+v, f(x)+g^T v) \mid v \in \mathbb{R}^d\}$ at $(x, f(x))$.

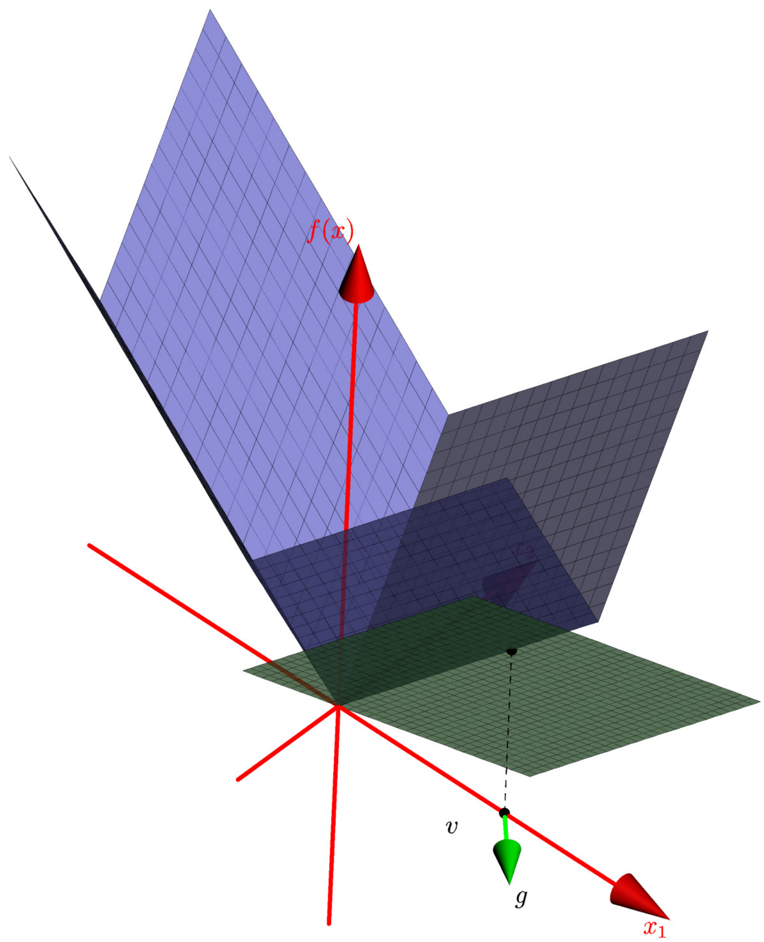
Consider $f(x) = |x|$ depicted below with some underestimating linear approximations.



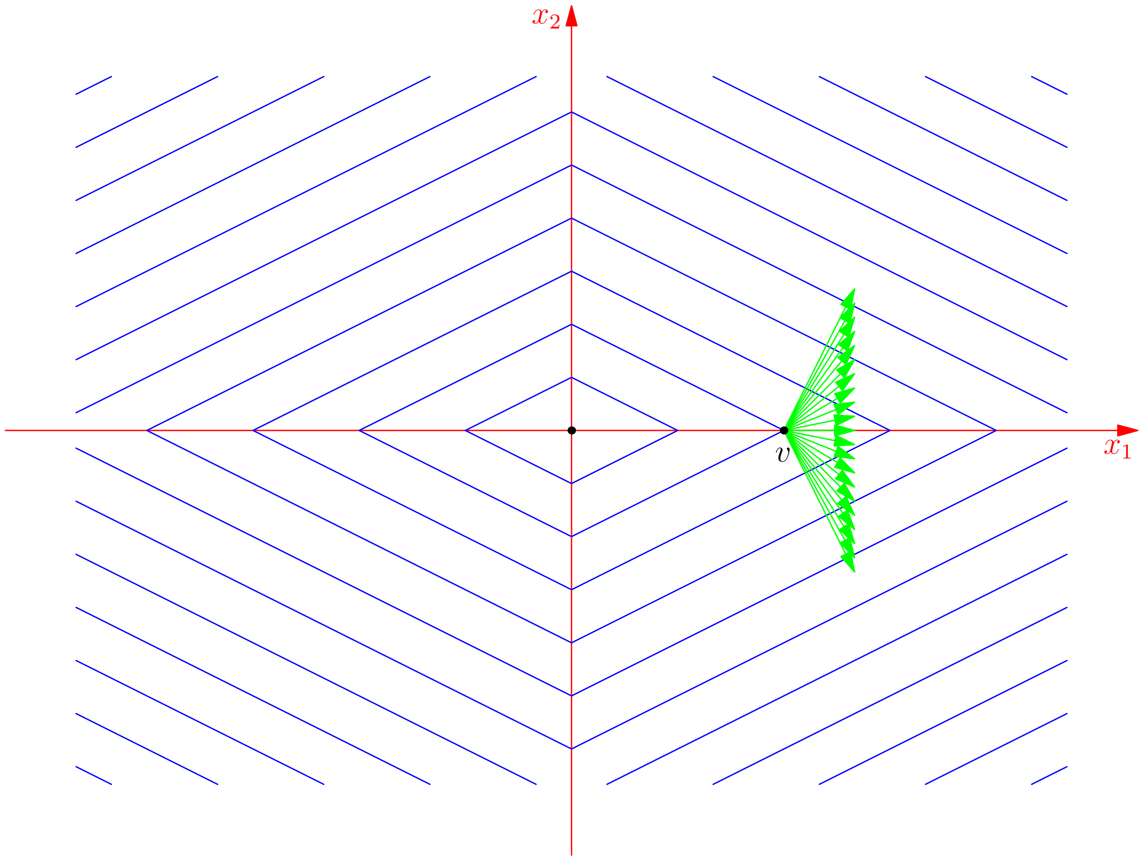
For $x \neq 0$ we have $\partial f(x) = \text{sgn}(x)$ since the function is convex and differentiable. At $x = 0$ we have $\partial f(x) = [-1, 1]$ since any slope between -1 and 1 will give an underestimating line. Note that the subgradients are **numbers** here since $f : \mathbb{R} \to \mathbb{R}$. Next we compute $\partial f(3, 0)$ where $f(x_1, x_2) = |x_1| + 2|x_2|$. The first coordinate of any subgradient must be 1 due to the $|x_1|$ part. The second coordinate can have any value between -2 and 2 to keep the hyperplane under the function.



$f(x)$
x_1


$f(x)$
v
g
x_1

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



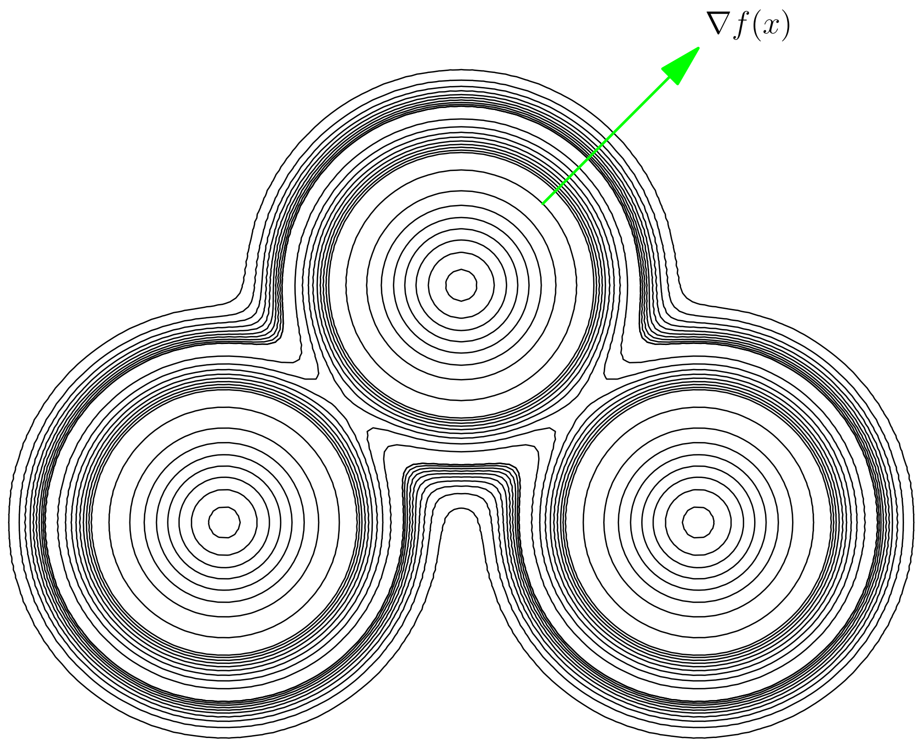
Contour Lines and Descent Directions

We can also look at the relationship between gradients and contour lines. Remember that for a function $f : \mathbb{R}^d \to \mathbb{R}$, the graph lies in $\mathbb{R}^{d+1}$ but the contour plot, level sets, gradients, and subgradients all live in $\mathbb{R}^d$. This is often a point of confusion. If $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and $x_0 \in \mathbb{R}^d$ with $\nabla f(x_0) \neq 0$ then $\nabla f(x_0)$ is normal to the level set $S = \{x \in \mathbb{R}^d \mid f(x) = f(x_0)\}$.

Proof sketch. Let $\gamma : (-1,1) \to S$ be differentiable path lying in S with $\gamma(0) = x_0$ (think of γ as describing a particle moving along the contour S). Then $f(\gamma(t)) = f(x_0)$ for all $t \in (-1,1)$ so that $\frac{d}{dt} f(\gamma(t)) = 0$. Thus we have

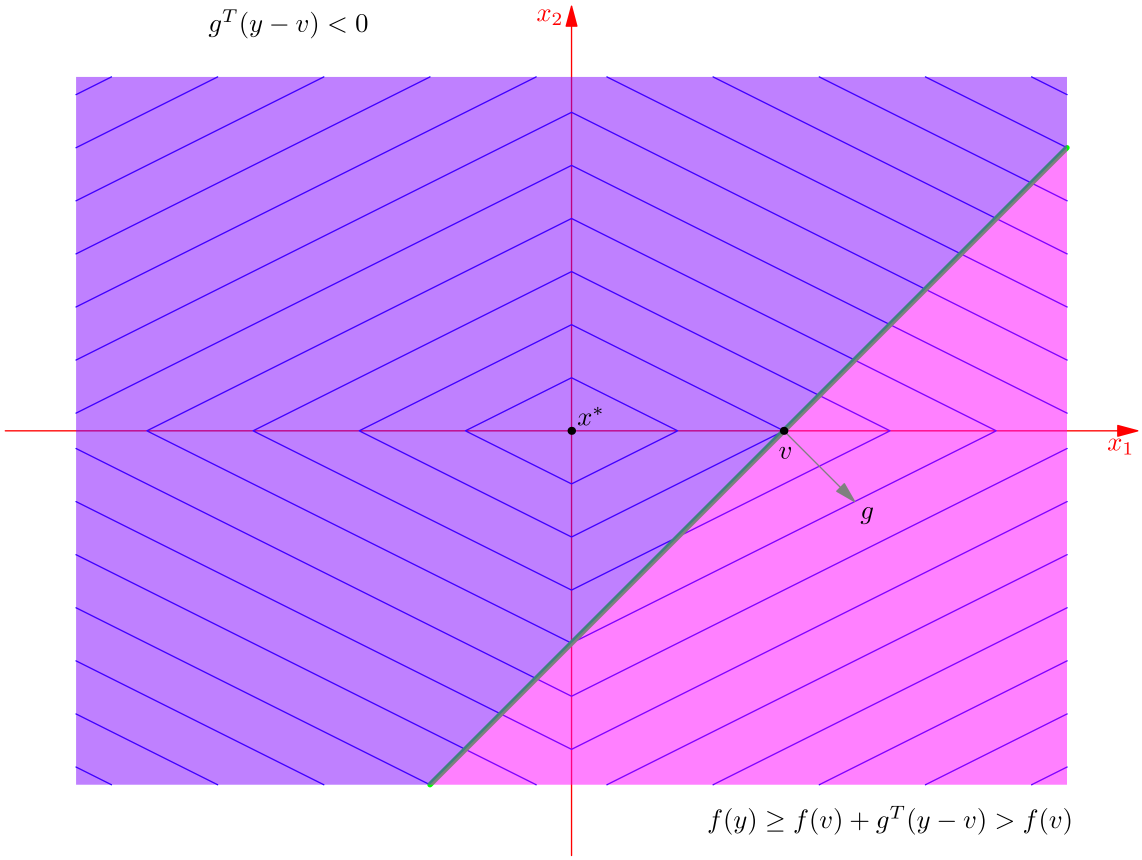
$$0 = \frac{d}{dt} f(\gamma(0)) = \nabla f(x_0)^T \gamma'(0),$$

so $\nabla f(x_0)$ is orthogonal to $\gamma'(0)$ (i.e., the gradient is orthogonal to the velocity vector of the particle γ that is tangent to the contour line at x_0). As γ is arbitrary, the result follows.



Now let's handle the non-differentiable case. Let $f : \mathbb{R}^d \to \mathbb{R}$ have subgradient g at x_0. The hyperplane H orthogonal to g at x_0 must *support* the level set $S = \{x \in \mathbb{R}^d \mid f(x) = f(x_0)\}$. That is, H passes through x_0 and all of S lies on one side of H (the side containing $-g$). This is immediate since any point y lying strictly on the side containing g must have

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



Even though points on the g side of H have larger f-values than $f(x_0)$, it is not true that points on the $-g$ side have smaller f-values. In other words, if g is a subgradient it may be true that $-g$ is not a descent direction (this is the case above). Using the same logic we obtain the following theorem.

Theorem 4. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex, let $x_0 \in \mathbb{R}^d$ not be a minimizer, let g be a subgradient of f at x_0, and suppose $x_* \in \mathbb{R}^d$ is a minimizer of f. Then for sufficiently small $t > 0$*

$$\|x_* - (x_0 - tg)\|_2 < \|x_* - x_0\|_2.$$

In other words, stepping in the direction of a negative subgradient brings us closer to a minimizer.

In fact, we can just choose t in the interval

$$t \in \left(0, \frac{2(f(x_0) - f(x^*))}{\|g\|_2^2}\right),$$

but since we usually don't know $f(x^*)$ this is of limited use.

This theorem suggests the following algorithm called Subgradient Descent.

1. Let $x^{(0)}$ denote the initial point.

2. For $k = 1, 2, \ldots$

    (a) Assign $x^{(k)} = x^{(k-1)} - \alpha_k g$, where $g \in \partial f(x^{(k-1)})$ and α_k is the step size.

    (b) Set $f_{\text{best}}^{(k)} = \min_{i=1,\ldots,k} f(x^{(i)})$. (Used since this isn't a descent method.)

Unfortuntely, there aren't any good stopping conditions worth mentioning. Recall that f is called Lipschitz with constant L if

$$|f(x) - f(y)| \leq L\|x - y\|$$

for all x, y.

Theorem 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and Lipschitz with constant G, and let x^* be a minimizer. For a fixed step size t, the subgradient method satisfies:*

$$\lim_{k\to\infty} f(x_{best}^{(k)}) \leq f(x^*) + G^2 t/2.$$

For step sizes respecting the Robbins-Monro conditions,

$$\lim_{k\to\infty} f(x_{best}^{(k)}) = f(x^*).$$

Subgradient descent can be fairly slow, with a provable convergence rate of $O(1/\epsilon^2)$ to achieve an error of order ϵ. Recall that the nice case for (unaccelerated) gradient descent was $O(1/\epsilon)$.