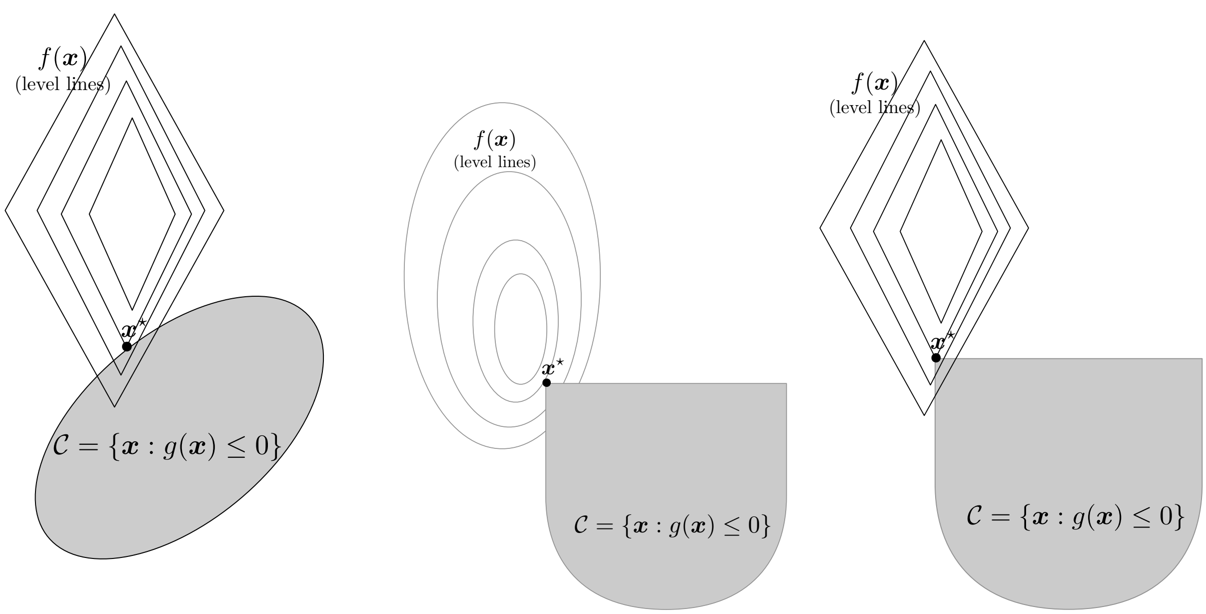
Nonsmooth constrained optimization

Just as with unconstrained optimization, we can generalize the theory of constrained optimization to the cases where the functions involved are not differentiable. Ultimately, this amounts to using the subdifferential in place of the gradient, but the analysis is more complicated — fully establishing necessary and sufficient conditions for optimality requires a few concepts from convex analysis that we have not covered in this course. The purpose of this section is to state what these results are, and draw parallels to the smooth case; we will forgo the careful mathematical arguments we gave in the previous sections. The classic references for this type of material include [Roc70, RW98, HUL01].

"Nonsmooth" can mean that f is not differentiable, the g_m are not differentiable, or both:



f nonsmooth, g smooth | f smooth, g nonsmooth | f nonsmooth, g nonsmooth

Hyperplanes and halfspaces

Hyperplanes and halfspaces are both very simple constructs, but they are crucial to how we approach non-smooth constrained optimization so it is worth defining them more carefully.

A **hyperplane** is an affine set of dimension $N-1$; it has the form

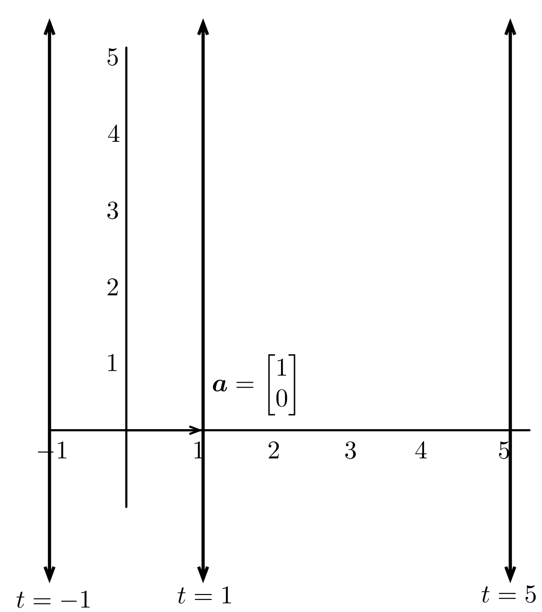
$$\{\boldsymbol{x} \in \mathbb{R}^N \ : \ \langle \boldsymbol{x}, \boldsymbol{a} \rangle = t\}$$

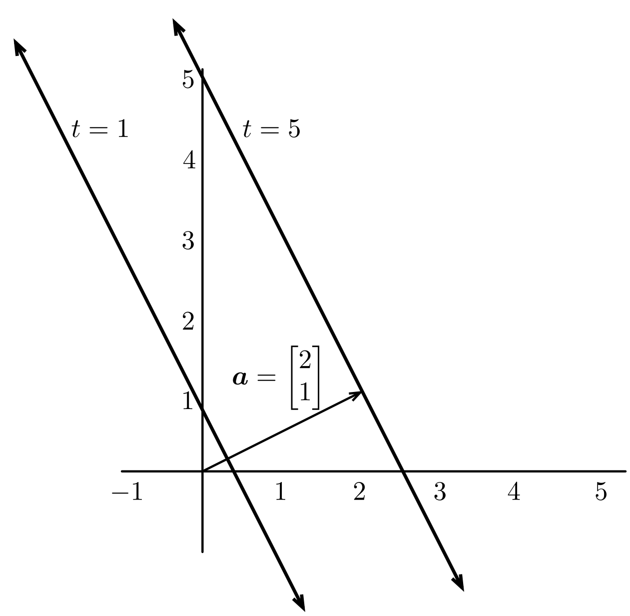
for some fixed vector $\boldsymbol{a} \neq \mathbf{0}$ and scalar t. When $t = 0$, this set is a subspace of dimension $N-1$, and contains all vectors that are orthogonal to $\boldsymbol{a}$. For $t \neq 0$, this is an affine space consisting of all the vectors orthogonal to $\boldsymbol{a}$ (call this set $\mathcal{A}^\perp$) offset to some $\boldsymbol{x}_0$:

$$\{\boldsymbol{x} \in \mathbb{R}^N \ : \ \langle \boldsymbol{x}, \boldsymbol{a} \rangle = t\} = \{\boldsymbol{x} \in \mathbb{R}^N \ : \ \boldsymbol{x} = \boldsymbol{x}_0 + \mathcal{A}^\perp\},$$

for any $\boldsymbol{x}_0$ with $\langle \boldsymbol{x}_0, \boldsymbol{a} \rangle = t$. We might take $\boldsymbol{x}_0 = t \cdot \boldsymbol{a}/\|\boldsymbol{a}\|_2^2$, for instance. The point is, $\boldsymbol{a}$ is a **normal vector** of the set.

Here are some examples in $\mathbb{R}^2$:

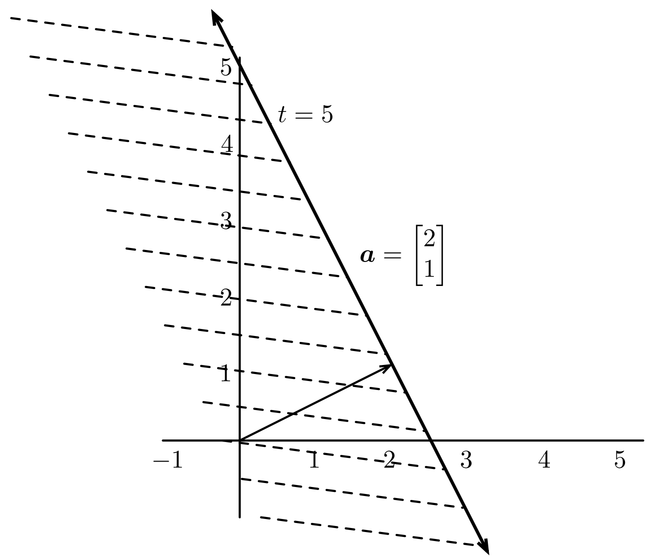




A **halfspace** is a set of the form

$$\{\boldsymbol{x} \in \mathbb{R}^N \ : \ \langle \boldsymbol{x}, \boldsymbol{a} \rangle \leq t\}$$

for some fixed vector $\boldsymbol{a} \neq \mathbf{0}$ and scalar t. For $t = 0$, the halfspace contains all vectors whose inner product with $\boldsymbol{a}$ is negative (i.e. the angle between $\boldsymbol{x}$ and $\boldsymbol{a}$ is greater than $90°$). Here is a simple example:



Supporting hyperplanes

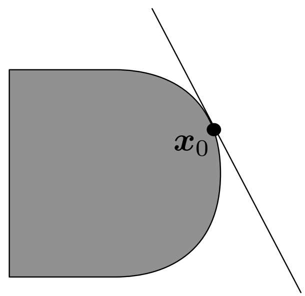
To aid our discussion, we introduce the following concept[1]: if $\boldsymbol{a} \neq \mathbf{0}$ satisfies $\langle \boldsymbol{x}, \boldsymbol{a} \rangle \leq \langle \boldsymbol{x}_0, \boldsymbol{a} \rangle$ for all $\boldsymbol{x} \in \mathcal{C}$, then

$$\mathcal{H} = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{a} \rangle = \langle \boldsymbol{x}_0, \boldsymbol{a} \rangle\}$$

is called a **supporting hyperplane** to $\mathcal{C}$ at $\boldsymbol{x}_0$.

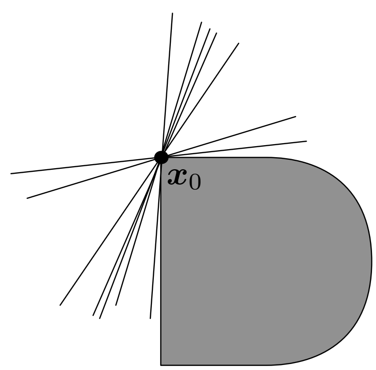
[1]We will assume throughout that $\mathcal{C}$ is a closed convex set.

Here's a picture:



The hyperplane is tangent to $\mathcal{C}$ at $\boldsymbol{x}_0$, and the halfspace $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{a} \rangle \leq \langle \boldsymbol{x}_0, \boldsymbol{a} \rangle\}$ contains $\mathcal{C}$.

Note that there might be more than one supporting hyperplane at an boundary point:

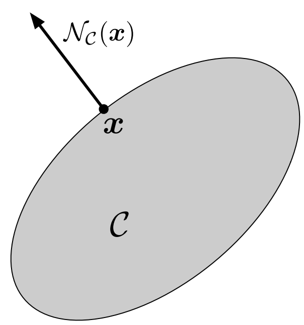


Normal cones

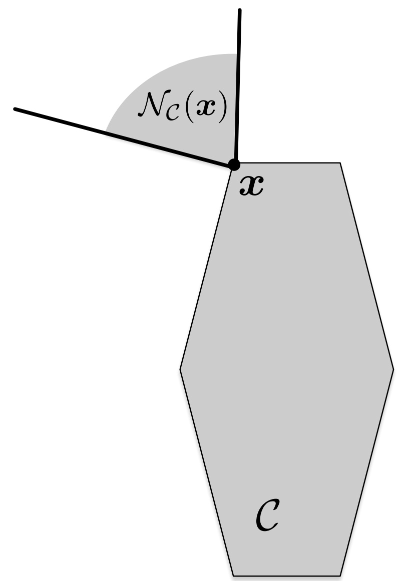
From The **normal cone** of $\mathcal{C}$ at a point $\boldsymbol{x}$ is defined as

$$\mathcal{N}_{\mathcal{C}}(\boldsymbol{x}) = \{\boldsymbol{u} \ : \ \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle \leq 0 \ \text{ for all } \boldsymbol{y} \in \mathcal{C}\}.$$

This is closely related to the set of all supporting hyperplanes of $\mathcal{C}$ that pass through $\boldsymbol{x}$. Some pictures will help at this point:



boundary of $\mathcal{C}$ smooth at $\boldsymbol{x}$



boundary of $\mathcal{C}$ nonsmooth at $\boldsymbol{x}$

Facts about $\mathcal{N}_{\mathcal{C}}(\boldsymbol{x})$ (stated without proof for now):

1. $\mathcal{N}_{\mathcal{C}}(\boldsymbol{x})$ is a closed convex cone
2. $\mathcal{N}_{\mathcal{C}} = \{\mathbf{0}\}$ if and only if $\boldsymbol{x} \in \operatorname{int}(\mathcal{C})$
3. If $\boldsymbol{u} \in \mathcal{N}_{\mathcal{C}}(\boldsymbol{x})$, then
$$\mathcal{H} = \{\boldsymbol{y} \ : \ \langle \boldsymbol{y}, \boldsymbol{u} \rangle = \langle \boldsymbol{x}, \boldsymbol{u} \rangle\},$$
is a supporting hyperplane of $\mathcal{C}$ at $\boldsymbol{x}$. The converse is also true. (This follows immediately from the definitions of supporting hyperplane and normal cone.)

Normal cones of sublevel sets. Take

$$\mathcal{C} = \{\boldsymbol{x} \ : \ g(\boldsymbol{x}) \leq 0\}, \quad g \text{ convex.}$$

Then for boundary points $\boldsymbol{x}$ (so $g(\boldsymbol{x}) = 0$),

1. If g is differentiable at $\boldsymbol{x}$, then

   $$\mathcal{N}_{\mathcal{C}}(\boldsymbol{x}) = \left\{\lambda \nabla g(\boldsymbol{x}),\ \lambda \geq 0\right\}.$$

   That is, the normal cone is the set of all vectors positively aligned with the gradient.

2. More generally (when g is not necessarily differentiable at $\boldsymbol{x}$),

   $$\mathcal{N}_{\mathcal{C}}(\boldsymbol{x}) = \left\{\lambda \partial g(\boldsymbol{x}),\ \lambda \geq 0\right\}.$$

   That is, the normal cone consists of all positive scalings of all the vectors in the subgradient.

Again, we are just stating these facts without proof for now.

Normal cones and indicator functions. Recall the definition of an indicator function of a set $\mathcal{C}$:

$$\iota_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \mathcal{C} \\ +\infty, & \boldsymbol{x} \notin \mathcal{C}. \end{cases}$$

We have that the subdifferential for $\iota_{\mathcal{C}}$ is the same as the normal cone,

$$\partial \iota_{\mathcal{C}}(\boldsymbol{x}) = \mathcal{N}_{\mathcal{C}}(\boldsymbol{x}), \quad \text{for } \boldsymbol{x} \in \mathcal{C}.$$

This is easy to see, as if $\boldsymbol{u} \in \partial \iota_{\mathcal{C}}(\boldsymbol{x})$, then

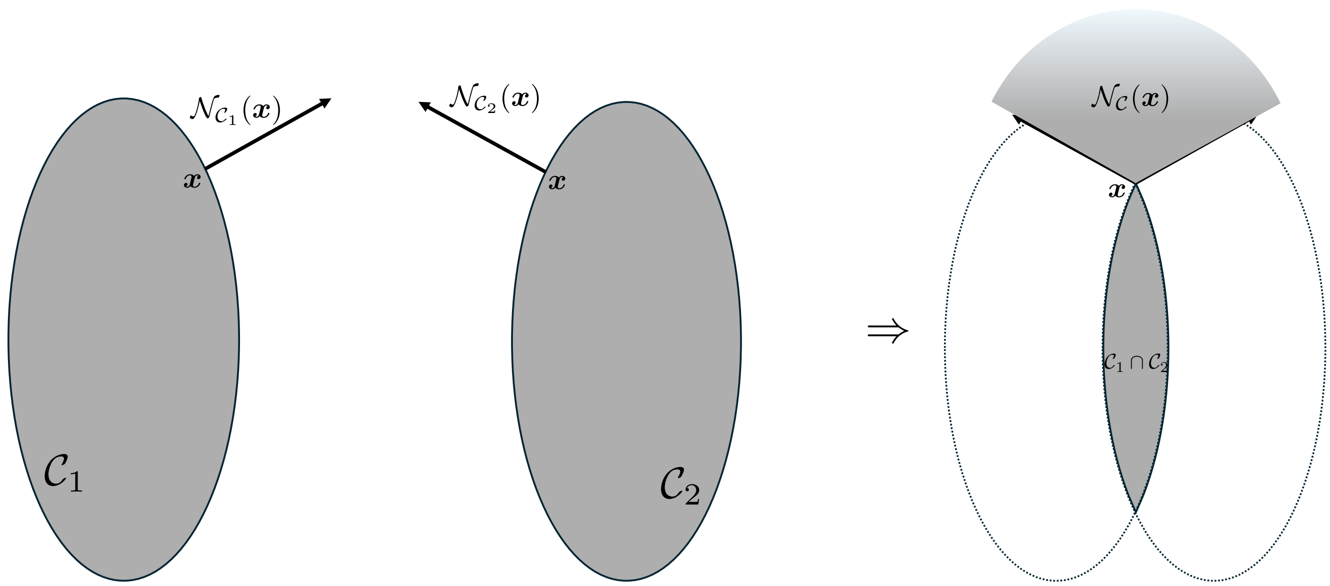
$$\iota_{\mathcal{C}}(\boldsymbol{y}) \geq \iota_{\mathcal{C}}(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle, \quad \text{for all } \boldsymbol{y} \in \mathcal{C},$$

which directly means $\langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle \leq 0$, since $\iota_{\mathcal{C}}(\boldsymbol{y}) = \iota_{\mathcal{C}}(\boldsymbol{x}) = 0$.

Normal cones of intersections of convex sets. Let $\mathcal{C}_1, \mathcal{C}_2$ be closed convex sets and take $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$. Then

$$\mathcal{N}_{\mathcal{C}}(\boldsymbol{x}) = \mathcal{N}_{\mathcal{C}_1}(\boldsymbol{x}) + \mathcal{N}_{\mathcal{C}_2}(\boldsymbol{x}).$$

That is, every $\boldsymbol{u} \in \mathcal{N}_{\mathcal{C}}(\boldsymbol{x})$ is generated by adding every vector from $\mathcal{N}_{\mathcal{C}_1}(\boldsymbol{x})$ to every vector from $\mathcal{N}_{\mathcal{C}_2}(\boldsymbol{x})$. Here is a representative picture:



You can establish this by using the relation between the subdifferential of indicator sets and the normal cone. Just note that

$$\iota_{\mathcal{C}_1 \cap \mathcal{C}_2}(\boldsymbol{x}) = \iota_{\mathcal{C}_1}(\boldsymbol{x}) + \iota_{\mathcal{C}_2}(\boldsymbol{x}),$$

and then take the subdifferential of both sides.

Optimality conditions for nonsmooth constrained optimization

We have already seen that if f is differentiable, then $\boldsymbol{x}^\star$ is a solution to

$$\underset{\boldsymbol{x}\in\mathcal{C}}{\text{minimize}}\ f(\boldsymbol{x})$$

if and only if $\boldsymbol{x}^\star \in \mathcal{C}$ and

$$-\nabla f(\boldsymbol{x}^\star) \in \mathcal{N}_\mathcal{C}(\boldsymbol{x}^\star) \quad \Leftrightarrow \quad \mathbf{0} \in \nabla f(\boldsymbol{x}^\star) + \mathcal{N}_\mathcal{C}(\boldsymbol{x}^\star).$$

For nonsmooth f, this condition becomes

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star) + \mathcal{N}_\mathcal{C}(\boldsymbol{x}^\star). \tag{1}$$

That is, there is a vector $\boldsymbol{u} \in \partial f(\boldsymbol{x}^\star)$ such that $-\boldsymbol{u} \in \mathcal{N}_\mathcal{C}(\boldsymbol{x}^\star)$.

Given what we have done so far, we can establish (1) quickly as follows. The constrained optimization problem above is equivalent to the unconstrained problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\ f(\boldsymbol{x}) + \iota_\mathcal{C}(\boldsymbol{x}),$$

meaning that $\boldsymbol{x}^\star$ is a solution if and only if $\boldsymbol{x}^\star \in \mathcal{C}$ and

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star) + \partial \iota_\mathcal{C}(\boldsymbol{x}) = \partial f(\boldsymbol{x}^\star) + \mathcal{N}_\mathcal{C}(\boldsymbol{x}).$$

Similarly, we can use the properties above to derive analogous KKT conditions. We consider again the problem

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) \quad \text{subject to} \quad g_m(\boldsymbol{x}) \le 0, \quad m = 1, \ldots, M.$$

As before, the constraint set can be written $\mathcal{C} = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_M$, where

$$\mathcal{C}_m = \{\boldsymbol{x} \ :\ g_m(\boldsymbol{x}) \le 0\},$$

and so

$$\mathcal{N}_{\mathcal{C}}(\boldsymbol{x}) = \sum_{m=1}^{M} \mathcal{N}_{\mathcal{C}_m}(\boldsymbol{x}).$$

Thus $\boldsymbol{x}^\star$ is a solution if and only if $\boldsymbol{x}^\star \in \mathcal{C}$ and

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \mathcal{N}_{\mathcal{C}_m}(\boldsymbol{x}^\star). \tag{2}$$

Using the relation between $\mathcal{N}_{\mathcal{C}_m}(\boldsymbol{x})$ and $\partial g_m(\boldsymbol{x})$, we have the KKT conditions

$$\begin{aligned} g_m(\boldsymbol{x}^\star) &\leq 0, \quad m = 1, \ldots, M \\ \lambda_m^\star &\geq 0, \quad m = 1, \ldots, M \\ \lambda_m^\star g_m(\boldsymbol{x}^\star) &= 0, \quad m = 1, \ldots, M \end{aligned}$$

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m^\star \partial g_m(\boldsymbol{x}^\star). \tag{3}$$

It is easy to see that these are sufficient; if they hold, then we can make (2) hold by taking $\lambda_m^\star \partial g_m(\boldsymbol{x}^\star) \in \mathcal{N}_{\mathcal{C}_m}(\boldsymbol{x}^\star)$ when $\lambda_m^\star > 0$ (and so $g_m(\boldsymbol{x}^\star) = 0$) and taking $\mathbf{0} \in \mathcal{N}_{\mathcal{C}_m}(\boldsymbol{x}^\star)$ when $\lambda_m^\star = 0$. Necessity is a little harder to argue, but does end up holding under the same Slater conditions as in the differentiable case.

References

[HUL01] J-B. Hiriart-Urrut and C. Lemarechal. *Fundamentals of Convex Analysis*. Springer, 2001.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[RW98] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Springer, 1998.