Nonsmooth optimization

Most of the theory and algorithms that we have explored for convex optimization have assumed that the functions involved are differentiable — that is, smooth.

This is not always the case in interesting applications. In fact, nonsmooth functions can arise quite naturally in applications. We already have looked at optimization programs involving the hinge loss $\max(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} + b, 0)$, the ℓ_1 norm, the ℓ_∞ norm, and the nuclear norm — none of these is differentiable. As another example, suppose $f_1, \ldots, f_Q$ are all perfectly smooth convex functions. Then the pointwise maximum

$$f(\boldsymbol{x}) = \max_{1 \leq q \leq Q} \; f_q(\boldsymbol{x})$$

is in general not smooth.

[Picture]:

Fortunately, the theory for nonsmooth optimization is not too different than for smooth optimization. We really just need one new concept: that of a subgradient.

Subgradients

If you look back through the notes so far, you will see that the vast majority of the time we use the gradient of a convex function, it is in the context of the inequality

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}), \quad \text{for all } \boldsymbol{y} \in \operatorname{dom} f.$$

[Picture]:

This is a very special property of convex functions, and it led to all kinds of beautiful results.

When convex f is not differentiable at a point $\boldsymbol{x}$, we can more or less reproduce the entire theory using subgradients. A *subgradient* of f at $\boldsymbol{x}$ is a vector $\boldsymbol{g}$ such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}), \quad \text{for all } \boldsymbol{y} \in \operatorname{dom} f.$$

Unlike gradients for smooth functions, there can be more than one subgradient of a nonsmooth function at a point. We call the collection of subgradients the *subdifferential* at $\boldsymbol{x}$:

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{g} \;:\; f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}), \quad \text{for all } \boldsymbol{y} \in \operatorname{dom} f\}.$$

[Picture]:

Facts:

1. If f is convex and differentiable at $\boldsymbol{x}$, then the subdifferential contains exactly one vector: the gradient,
$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}.$$

2. If f is convex on dom f, then the subdifferential is non-empty and bounded at all $\boldsymbol{x}$ in the interior of dom f.

For non-convex f, these two points do not hold in general. The gradient at a point is not necessarily a subgradient
[Picture]

and there can also be points where neither the gradient nor subgradient exist, e.g. $f(x) = -\sqrt{|x|}$ for $x \in \mathbb{R}$
[Picture]

Geometrically, if $\boldsymbol{g}$ is a subgradient at $\boldsymbol{x}$, then

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) \quad \Rightarrow \quad \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \leq 0,$$

i.e. a non-zero subgradient at $\boldsymbol{x}$ defines a supporting hyperplane of the sublevel set $\{\boldsymbol{y} \ : \ f(\boldsymbol{y}) \leq \ f(\boldsymbol{x})\}$.
[Picture]

Optimality conditions for unconstrained optimization

(New and Improved!!)

With the right definition in place, it is very easy to re-derive the central mathematical results in this course for general[1] convex functions.

Let $f(\boldsymbol{x})$ be a general convex function. Then $\boldsymbol{x}^\star$ is a solution to the unconstrained problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x})$$

if and only if

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star).$$

Proof of this statement is so easy it's stupid:
Suppose $\mathbf{0} \in \partial f(\boldsymbol{x}^\star)$. Then

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}^\star) + \mathbf{0}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \\ &= f(\boldsymbol{x}^\star) \end{aligned}$$

for all $\boldsymbol{y} \in \operatorname{dom} f$. Thus $\boldsymbol{x}^\star$ is optimal. Likewise, if $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^\star)$ for all $\boldsymbol{y} \in \operatorname{dom} f$, then of course it must also be true that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^\star) + \mathbf{0}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{y}$, and so $\mathbf{0} \in \partial f(\boldsymbol{x}^\star)$.

[1] Meaning not necessarily differentiable.

Example: the LASSO

Consider the ℓ_1 regularized least-squares problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\tau\|\boldsymbol{x}\|_1.$$

The ℓ_1 norm is not differentiable at any $\boldsymbol{x}$ that has at least one coordinate equal to zero. We have also seen that solutions to this program tend to be sparse, that is, have many coordinates that are indeed equal to zero. So the nonsmoothness is kicking in at exactly the points we are interested in.

We can quickly translate the general result $\mathbf{0}\in\partial f(\boldsymbol{x}^\star)$ into a useful set of optimality conditions. We need to compute the subdifferential of $f(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\tau\|\boldsymbol{x}\|_1$. The first term is smooth, so the subdifferential just contains the gradient:

$$\partial f(\boldsymbol{x})=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y})+\tau\partial\|\boldsymbol{x}\|_1,$$

where as we have seen already

$$\partial\|\boldsymbol{x}\|_1=\left\{\boldsymbol{u}\ :\ u[i]=\operatorname{sign}x[i],\ i\in\Gamma(\boldsymbol{x});\ |u[i]|\leq 1,\ i\notin\Gamma\right\},$$

where $\Gamma(\boldsymbol{x})$ is the set of indexes where $\boldsymbol{x}$ is non-zero:

$$\Gamma(\boldsymbol{x})=\{i\ :\ x[i]\neq 0\}.$$

Thus the optimality condition

$$\mathbf{0}\in\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}^\star-\boldsymbol{y})+\tau\partial\|\boldsymbol{x}^\star\|_1,$$

means that $\boldsymbol{x}^\star$ is optimal if and only if

$$\begin{aligned}\boldsymbol{a}_i^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}^\star)&=\tau\operatorname{sign}x[i], && i\in\Gamma(\boldsymbol{x}^\star),\\ |\boldsymbol{a}_i^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}^\star)|&\leq\tau, && i\notin\Gamma(\boldsymbol{x}^\star),\end{aligned}$$

where $\boldsymbol{a}_i^{\mathrm{T}}$ is the ith column of $\boldsymbol{A}$.

Homework for the industrious student

1. Show that the subdifferential for a convex function at a point $\boldsymbol{x}$ is always a convex set.

2. Consider the piecewise linear function
$$f(\boldsymbol{x}) = \max_{m=1,\ldots,M} \left(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} + b_m\right).$$
Show that
$$\partial f(\boldsymbol{x}) = \operatorname{conv}\left(\{\boldsymbol{a}_m \ : \ m \in \Gamma(\boldsymbol{x})\}\right),$$
where $\Gamma(\boldsymbol{x})$ are the indexes of the "active" functions in the max at $\boldsymbol{x}$:
$$\Gamma(\boldsymbol{x}) = \{m \ : \ f_m(\boldsymbol{x}) = f(\boldsymbol{x})\},$$
and $\operatorname{conv}(\mathcal{X})$ is the convex hull of the set $\mathcal{X}$.

3. Consider the optimization problem
$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{x}), \quad \text{where} \quad f(\boldsymbol{x}) = \max_{m=1,\ldots,M} \left(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} + b_m\right).$$
Show that the optimality condition $\mathbf{0} \in \partial f(\boldsymbol{x}^\star)$ is the same as saying that there exists a $\boldsymbol{\lambda}$ such that
$$\boldsymbol{\lambda} \geq \mathbf{0}, \ \mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda} = 1, \ \sum_{m=1}^{M} \lambda_m \boldsymbol{a}_m = \mathbf{0}, \quad \text{and } \lambda_m \neq 0 \text{ for } m \notin \Gamma(\boldsymbol{x}^\star).$$

4. Show that the conditions above are exactly the same as the KKT conditions for the linear program
$$\underset{t\in\mathbb{R},\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ t \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \leq t\mathbf{1}$$

Optimality conditions for constrained optimization

The theory for constrained optimization can also be stated in terms of subgradients. We will state the main results here without proof, as the proofs are almost exactly the same as in the smooth case.

Let f be a convex function that we want to minimize over a convex set $\mathcal{C}$:

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

Then $\boldsymbol{x}^\star$ is a minimizer of the program above if and only if $\boldsymbol{x}^\star \in \mathcal{C}$ and there exists a $\boldsymbol{g} \in \partial f(\boldsymbol{x}^\star)$ such that

$$\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}^\star) \geq 0 \quad \text{for all} \quad \boldsymbol{y} \in \mathcal{C}.$$

[Picture:]

We have already seen that for $\boldsymbol{x}$ such that $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^\star)$ we have $\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^\star) \leq 0$. This means that condition above amounts to the existence of a *separating hyperplane* between the constraint set $\mathcal{C}$ and the sublevel set $\{\boldsymbol{x} \;:\; f(\boldsymbol{x}) \leq f(\boldsymbol{x}^\star)\}$.

If the constraints are in functional form, we can also generalize the KKT conditions.

Let $f_0, \ldots, f_M$ be (not necessarily differentiable) convex functions, and consider the optimization program

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f_0(\boldsymbol{x}) \quad \text{subject to} \quad f_m(\boldsymbol{x}) \leq 0, \quad m = 1, \ldots, M.$$

If strong duality holds, the $\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star$ are primal,dual optimal if and only if

$$\begin{array}{ll} \text{(K1)} & f_m(\boldsymbol{x}^\star) \leq 0, \quad m = 1, \ldots, M \\ \text{(K2)} & \boldsymbol{\lambda}^\star \geq \mathbf{0} \\ \text{(K3)} & \lambda_m^\star f_m(\boldsymbol{x}^\star) = 0, \quad \text{for } m = 1, \ldots, M \\ \text{(K4)} & \mathbf{0} \in \partial f_0(\boldsymbol{x}^\star) + \displaystyle\sum_{m=1}^{M} \lambda_m^\star \partial f_m(\boldsymbol{x}^\star) \end{array}$$

As before, the last condition can be interpreted in terms of the Lagrangian

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f_0(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{x}).$$

It says that

$$\mathbf{0} \in \partial_{\boldsymbol{x}} L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star),$$

and so $\boldsymbol{x}^\star$ must be the minimizer to the unconstrained program

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\ L(\boldsymbol{x}, \boldsymbol{\lambda}^\star).$$

Technical Notes: Set scaling and addition

While the gradient of a smooth function at a point is a single vector, the subdifferential is a *set* of vectors. In the notes above, we have multiplied subdifferentials by scalars, added vectors to them, and added multiple subdifferentials together. We should make it clear what these operations mean.

Let $\mathcal{X} \subset \mathbb{R}^N$ be an arbitrary set of vectors. Then we have the following definitions:

1. **Scalar multiplication**. For any $a \in \mathbb{R}$, we define
$$a\mathcal{X} = \{a \cdot \boldsymbol{x} \ : \ \boldsymbol{x} \in \mathcal{X}\}.$$
2. **Vector addition**. For any $\boldsymbol{v} \in \mathbb{R}^N$, we define
$$\mathcal{X} + \boldsymbol{v} = \{\boldsymbol{x} + \boldsymbol{v} \ : \ \boldsymbol{x} \in \mathcal{X}\}.$$
3. **Set addition**. For another $\mathcal{Y} \subset \mathbb{R}^N$, we define
$$\mathcal{X} + \mathcal{Y} = \{\boldsymbol{x} + \boldsymbol{y} \ : \ \boldsymbol{x} \in \mathcal{X}, \ \boldsymbol{y} \in \mathcal{Y}\}.$$

Exercise: Let $\mathcal{X} \subset \mathbb{R}^2$ be the unit ball for the ℓ_2 norm:
$$\mathcal{X} = \{\boldsymbol{x} \ : \ \|\boldsymbol{x}\|_2 \leq 1\}.$$

1. Sketch the sets $a\mathcal{X}$ for $a = 2, 1/2, -1/2$.
2. Sketch the set $\mathcal{X} + \boldsymbol{v}$ for $\boldsymbol{v} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$.
3. Sketch the set $\epsilon\mathcal{X} + \mathcal{Y}$, where
$$\mathcal{Y} = \{\boldsymbol{y} \ : \ \|\boldsymbol{y}\|_1 \leq 1\},$$
for $\epsilon = .1$ and $\epsilon = 1$.