1 Minimization Through Composition

Consider the Euclidean function. Despite being common in optimization, this function is neither smooth nor strongly convex. For this reason, it would be natural to assume that minimizing this norm via iterative methods would be a slowly converging process. However, if we instead minimize $f(x) = ||x||_2^2$, we have much more desirable properties. The function f is both strongly convex and smooth and maintains the same minimizer as the Euclidean norm. This was accomplished by composing our origin function with a monotonically increasing function. Can we apply a similar process to other convex functions to obtain reformulations with better properties? In this section, we explore smoothing convex functions through composition.

Let $f : X \to \mathbb{R}$ be a continuous, convex function over a compact set $X \subset \mathbb{R}^n$. For simplicity, let us assume that f has a unique minimizer x^*. Our goal is to find a function $g : \mathbb{R} \to \mathbb{R}$ such that $h(x) := (g \circ f)(x)$ is a convex and differentiable over X with

$$x^* := \operatorname*{argmin}_{x \in X} f(x) = \operatorname*{argmin}_{x \in X} h(x).$$

We know that to maintain the minimizer, we must enforce that $g'(f(x_0)) > 0$ for any $x_0 \in X \setminus \{x^*\}$. To study differentiability, we will utilize subdifferential sets. Recall that a convex function has a nonempty subdifferential set at any point in its domain. Furthermore, f is differentiable at $x_0 \in X$ if and only if the subdifferential of f at x_0 is a singleton. Let $x_0 \in X$. From convex analysis, it can be shown that

$$\partial h(x_0) = \{\alpha\beta \mid (\alpha, \beta) \in \partial g(f(x_0)) \times \partial f(x_0)\}.$$

We will proceed into two cases.

Assume that f is differentiable at $x_0 \in X$. Then $\partial f(x_0)$ is a singleton. Thus, in order for $\partial h(x_0)$ to be a singleton, either g must be differentiable at $f(x_0)$ or $\nabla f(x_0) = 0$. Since we do not want to assume knowledge of points satisfying $\nabla f(x_0) = 0$, we will instead enforce that g be differentiable everywhere.

Now consider the case when f is non-differentiable at $x_0 \in \mathbb{R}^n$. Then $|\partial f(x_0)| > 1$. Thus, in order for $\partial h(x_0)$ to be a singleton, we must have that $g'(f(x_0)) = 0$.

To summarize, to ensure that h is differentiable over X, we must choose a differentiable function g satisfying $g'(f(x_0)) = 0$ for any point $x_0 \in X$ such that f is non-differentiable at x_0. However, in order for h be convex and have the same minimizer as f, we must also require $g'(f(x_0)) > 0$ for $x_0 \in X \setminus \{x^*\}$. Here, we see the difficulties of choosing our function g. We cannot create smoothness at $x_0 \in X$ unless x_0 is itself the minimizer of f. This also explains the previously noted phenomenon with the Eulcidean norm. $g(x) = x^2$ is differentiable and convex over the image of $f(x) = ||x||_2^2$ and satisfies $g'(f(x^*)) = g'(0) = 0$. Unfortunately, for most functions, this will not be the case. We leave the analysis of the (strong) convexity of h to future study.

2 On the Duality Between Smoothing and Catalyst Algorithms

A popular direction in current literature is to approximate an objective function f with a separate function f_μ parameterized by $\mu > 0$ such that the minimizer of f_μ is close to the minimizer of f and f_μ has additional optimization properties. These additional properties will allow for accelerated methods to be used to minimize f_μ which are not possible with f. If the acceleration is large and the difference in minimizers is not, then minimizing f_μ may be a much better approach to finding an approximate minimizer for f. For example, if f is convex, nonsmooth, then smoothing it lets us move from the subgradient error rate ($\mathcal{O}(1/\sqrt{t})$) to a faster rate ($\mathcal{O}(1/t)$). If f is smooth and convex but not strongly convex, then we may look to for a strongly convex f_μ to benefit from strongly convex methods. Here we show that the two proposed ideas, smoothing and catalyst methods, are merely duals of each other. We will heavily rely on convex conjugate theory - which we will briefly review.

For any function real valued function f, we define its convex conjugate via

$$f^*(y) = \sup_{x \in \operatorname{dom} f} y^T x - f(x).$$

Furthermore, if f is proper, lower semicontinuous and convex, then f satisfies the biconjugacy property $f = (f^*)^*$. In this case, we can rewrite f as

$$f(x) = \sup_{y \in \operatorname{dom} f^*} x^T y - f^*(y).$$

The functions f and f^* are related in many ways. One that will be of importance to us is the following.

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper, continuous, closed, convex function. Then f is strongly convex with strong convexity parameter $\mu > 0$, if and only if f^* is differentiable with Lipschitz continuous gradient and Lipschitz constant $1/\mu$.

In short, under ideal conditions, the strong convexity of f or f^* implies the smoothness of the other and vice versa. This suggests a duality relationship between smoothing methods and catalyst methods. By catalyzing (i.e. to make strongly convex) the convex conjugate, we construct a smooth approximation to the original function. By smoothing the convex conjugate, we construct a strongly convex approximation to the original function.

We can use this duality relationship to more easily understand popular algorithms such as Nesterov's smoothing technique. By the above theorem, if we want to construct a smooth approximation f_μ to f, then it suffices to catalyze f^*. Using the standard catalyst approach, we can construct a strongly convex approximation to f^* via

$$f^*_\mu(y) = f^*(y) + \mu d(y) = \sup_{x \in \operatorname{dom} f} y^T x - f(x) + \mu d(y)$$

where $d : \mathbb{R}^n \to \mathbb{R}$ is a strongly convex prox function. Taking the convex conjugate again to get back f_μ from the biconjugacy property, we obtain

$$f_\mu(x) = \sup_{y \in \operatorname{dom} f^*} x^T y - f^*_\mu(y) = \sup_{y \in \operatorname{dom} f^*} x^T y - f^*(y) - \mu d(y)$$

which is precisely Nesterov's smoothing technique. This suggests another method of constructing catalyst/smoothing algorithms. For any given catalyst algorithm, simply applying it to the convex conjugate of a sufficiently nice function f will result in a smoothing algorithm for f. We can use this idea to construct another catalyst technique via Moreau-Yosida smoothing.

3 Proximal Problems

Iterative minimization techniques of a convex objective function can be understood and interpreted in many different ways. For example, if one were to try to minimize a differentiable function f, a very natural idea is a gradient descent method:

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left|\left|x - (x_k - \lambda \nabla f(x_k))\right|\right|^2 .$$

Here, we are constructing the point x_{k+1} by moving in the negative gradient direction and then projecting back onto some constraint set X. However, we can also rewrite the gradient descent step by expanding the norm term to obtain

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \langle \nabla f(x_k), x \rangle + \frac{1}{2\lambda} \left|\left|x - x_k\right|\right|^2 .$$

This formulation reveals to us a different interpretation of a gradient descent step - instead of minimizing f, we will iteratively minimize a linear approximation of f while also keeping our new iterative close to our previous one. The additional proximal term $\left|\left|x - x_k\right|\right|^2$ keeps us close to the previous iterate, at the cost of potentially making the subproblem not a linear one. The importance of each term is balanced by the parameter λ which we previously understood as a step size. Indeed, from this interpretation, if our step size is small, then it will be more important to keep close to the previous iteration.

From the latter interpretation of gradient descent, we can recover a method for minimizing nonsmooth functions. In the absence of an easy linear approximation to the function, we can instead iteratively solve a proximal problem, i.e.

$$x_{k+1} = \operatorname{prox}_{\lambda f}(x_k) := \operatorname*{argmin}_{x \in X} f(x) + \frac{1}{2\lambda} \left|\left|x - x_k\right|\right|^2 .$$

Here, the idea remains the same: we simply minimize our function while keeping close to a previously computed point. The two are balanced by the parameter λ. Unfortunately, unlike in the gradient descent step, the proximal problem is not computable via projection. Although this idea is straight forward in theory, computing a solution to the proximal problem may be no easier than minimizing f. In instances where a solution to the proximal problem can be obtained easily, the proximal point algorithm is clearly preferred, but this is not always the case.

4 Moreau Envelope and Simple Smoothing Argument

Define the infimal convolution operator applied to two functions f and g as

$$(f \Box g)(v) = \inf_x f(x) + g(v - x).$$

Now suppose that f is a nonsmooth function and let $g(x) = \frac{1}{2} \left|\left|x\right|\right|^2$. Then the Moreau envelope of f with constant $\lambda > 0$ is defined as

$$M_{\lambda,f}(v) := (\lambda f \Box g)(v) = \inf_x f(x) + \frac{1}{2\lambda} \left|\left|x - v\right|\right|^2 .$$

The Moreau envelope has many nice properties. It has domain in $\mathbb{R}^n$, is strongly convex and continuously differentiable, and has the same minimizers as f. However, the Moreau envelope is obviously harder to compute first order information for. Indeed, from above the zero-th order information is a minimization problem and it can be shown that

$$\nabla M_{\lambda,f}(v) = \frac{1}{\lambda}(v - \mathrm{prox}_{\lambda f}(v))$$

i.e, the gradient is computed via a proximal problem. Nonetheless, the Moreau envelope can be a powerful tool in the minimization of nonsmooth functions as it creates an auxillary problem to solve that has better properties than the underlying problem.

Let us briefly observe how the Moreau envelope achieves its smoothing. Letting f^* denote the convex conjugate of a real valued function f, it can be shown that

$$(f \Box g)^* = f^* + g^*.$$

Since $M_f^{**} = M_f$ and g is self-dual, we conclude that

$$M_f = (M_f^*)^* = (f^* + g)^*.$$

That is, the Moreau envelope can be obtained by catalyzing the convex conjugate of f, and then taking its conjugate again. Since the convex conjugate of a strongly convex function must be smooth, the Moreau envelope itself is then $(1/\lambda)$-smooth.