Terence Tao suggests keeping track of mathematical tricks that are difficult or unnecessary to remember. This note will blindly follow his advice. Below is a list of materials that I thought useful to write down. These include arguments I needed to formalize, small tricks I encountered during research, or even interesting problems suitable for a qualifying exam. I anticipate these to be organized very poorly at first, perhaps improving as the list becomes larger.

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 CndG Step Size

The procedure CndG is listed in [1] is as follows

---

procedure $u^+ = \text{CndG}(g, u, \beta, \eta)$

1. Set $u_1 = u$ and $t = 1$.
2. Let v_t be the optimal solution for the subproblem of
$$V_{g,u,\beta}(u_t) := \max_{x \in X} \langle g + \beta(u_t - u), u_t - x \rangle. \tag{3.1}$$
3. If $V_{g,u,\beta}(u_t) \leq \eta$, set $u^+ = u_t$ and terminate.
4. Set $u_{t+1} = (1 - \alpha_t)u_t + \alpha_t v_t$ where $\alpha_t = \min\{1, \langle \beta(u - u_t) - g, v_t - u_t \rangle / (\beta ||v_t - u_t||^2)\}$.
5. Set $t \leftarrow t + 1$ and go to step 2.

end procedure

This procedure can be viewed as the classical algorithm which minimizes a linear objective at each step applied to the function $\phi(x) = \langle g, x \rangle + \beta ||x - u||^2 / 2$. For instance, the maximizer of (3.1) is equivalent to the minimizer of $\langle \phi'(u_t), x \rangle$. The termination criteria is whenever $V_{g,u,\beta}(u_t)$ is smaller than the tolerance η. The stepsize optimal step size α_t is then the solution to
$$\alpha_t = \operatorname*{argmin}_{\alpha \in [0,1]} \phi((1 - \alpha)u_t + \alpha v_t). \tag{3.2}$$
We will show that $\alpha_t = \min\{1, \langle \beta(u - u_t) - g, v_t - u_t \rangle / (\beta ||v_t - u_t||^2)\}$.

Proof. We use the KKT conditions to solve (3.2). Let us introduce dual variables μ_1 and μ_2 to deal with the constraints $\alpha \leq 1$ and $0 \leq \alpha$ respectively. Our conditions are as follows:

- Primal Feasibility:
$$0 \leq \alpha \leq 1$$
- Dual Feasibility
$$\langle (v_t - u_t), (g + \beta((1 - \alpha)u_t + \alpha v_t - u)) \rangle + \mu_1 - \mu_2 = 0$$
and
$$\mu_1, \mu_2 \geq 0$$
- Complementary Slackness
$$\mu_1(\alpha - 1) = 0, \text{ and } \mu_2 \alpha = 0.$$

With 2 dual variables, we have 4 cases. For $\mu_1 \neq 0$ and $\mu_2 \neq 0$, we yield no solution since both $\alpha = 0$ and $\alpha = 1$ must hold. For $\mu_1 = 0$ and $\mu_2 \neq 0$, we must have $\alpha = 0$ and $\mu_2 = \langle (v_t - u_t), (g + \beta(u_t - u)) \rangle$. But by the definition of v_t, we know that $\langle \phi'(u_t), v_t \rangle \leq \langle \phi'(u_t), u_t \rangle$. Thus, $\mu_2 = \langle v_t - u_t, \phi'(u_t) \rangle < 0$ since $\mu_2 \neq 0$. Since this contradicts dual feasibility, this case gives no solutions.

For $\mu_1 \neq 0$ and $\mu_2 = 0$, we have $\alpha = 1$ and $\mu_1 = \langle (u_t - v_t), (g + \beta(u_t - u)) \rangle$. By a similar argument, it can be shown that $\mu_1 \geq 0$. Thus, $\alpha = 1$ is a solution if $\langle (u_t - v_t), (g + \beta(u_t - u)) \rangle \neq 0$. Finally, whenever $\mu_1 = \mu_2 = 0$, α must satisfy

$$\langle v_t - u_t, g + \beta((1 - \alpha)u_t + \alpha v_t - u)) \rangle = 0.$$

Splitting the LHS into two inner products, it follows that

$$\alpha\beta \left|\left|v_t - u_t\right|\right|^2 + \langle \beta(u_t - u) + g, v_t - u_t \rangle = 0$$

and thus $\alpha = \langle \beta(u - u_t) - g, v_t - u_t \rangle / (\beta \left|\left|v_t - u_t\right|\right|^2)$. Note that by our previous discussion, the numerator is always positive and consequently $\alpha \geq 0$.

In summary, there are two possible KKT points. If

$$\langle \beta(u - u_t) - g, v_t - u_t \rangle / (\beta \left|\left|v_t - u_t\right|\right|^2) \in [0, 1],$$

i.e. if it is less than 1, then this is our solution. Otherwise, since the solution exists and is unique, it must be that $\alpha_t = 1$. We can summarize this result as $\alpha_t = \min \{1, \langle \beta(u - u_t) - g, v_t - u_t \rangle / (\beta \left|\left|v_t - u_t\right|\right|^2)\}$.

□

References

[1] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.