Brief notes on composite optimization

John Duchi
March 19, 2021

1 Composite optimization

The composite optimization problem is to

$$\min_x f(x) := h(c(x)) \text{ subject to } x \in X,$$

where $X$ is a closed convex set, $c : \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable function, and $h : \mathbb{R}^m \to \mathbb{R}$ is a convex function. For $c(x) = (c_1(x), \ldots, c_m(x))$, we let

$$\nabla c(x) := [\nabla c_1(x) \ \nabla c_2(x) \ \cdots \ \nabla c_m(x)] \in \mathbb{R}^{n \times m}$$

denote the gradient, so that $c(x + \Delta) = c(x) + \nabla c(x)^T \Delta + o(\|\Delta\|)$.

The key in problem (1) is to identify some (mild) restrictions on $h, c$ to make it possible to approximate the objective with a convex function, then repeatedly solve convex optimization problems to (approximately) minimize the objective. The simplest and most common version of this is to assume that $c : \mathbb{R}^n \to \mathbb{R}^m$ has Lipschitz derivatives (w.r.t. a norm that may be problem specific), and $h : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous, meaning that for Lipschitz constants $L_c, L_h$ we have

$$\|\nabla c(x) - \nabla c(y)\|_a \leq L_c \|x - y\|_b,$$

where $a, b$ just indicate that these may be different norms, and $|h(x) - h(y)| \leq L_h \|x - y\|$, again for (what may be) a different norm $\|\cdot\|$. Optimization methods typically iterate by (1) making an approximation—or a model—to the objective $f$ to be optimized, and (2) minimizing this approximation, perhaps with regularization. In this note, I will use $f_x$ to denote the model of $f$ around the point $x$. Then a generic (Euclidean-type, for now) optimization method iterates via

$$x_{k+1} = \arg\min_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\},$$

where $\alpha_k > 0$ denotes a stepsize, which may be prespecified or adaptively chosen. The (projected) gradient method uses the linear model

$$f_x(y) := f(x) + \nabla f(x)^T (y - x),$$

Newton’s method uses the second order model

$$f_x(y) := f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x),$$

and the proximal-point method [2] uses no approximation at all, modeling $f_x(y) := f(y)$.  

In the case of a composite objective (1), a natural idea is to use a convex model, linearizing the internal component $c(x)$ while not touching $h$, yielding the model

$$f_x(y) := h(c(x) + \nabla c(x)^T(y - x)),$$

which is evidently convex in $y$. Thus we may iterate

$$x_{k+1} = \arg\min_{x \in X} \left\{ h(c(x_k) + \nabla c(x_k)^T(x_{k+1} - x_k)) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$  \hspace{1cm} (3)

Note that the iteration (3) is guaranteed to be a descent method (or at least, not ascend) if the stepsize $\alpha = \alpha_k$ is chosen in such a way that

$$|f_x(y) - f(y)| \leq \frac{1}{2\alpha} \|x - y\|^2,$$

so that the model $f_x$ is accurate enough for $f = h \circ c$ that the error is (at most) quadratic.

A common case for this approximation guarantee is when the function $c$ is smooth and $h$ itself is Lipschitz. In particular, assume that $c$ has $L$-Lipschitz gradient with respect to the $\ell_2$-norm, meaning that

$$\|\nabla c(x) - \nabla c(y)\|_2 \leq L \|x - y\|_2$$

(where the first norm is on matrices and is the $\ell_2$-operator norm), and that $h$ is $M$-Lipschitz, meaning that

$$|h(z_0) - h(z_1)| \leq M \|z_0 - z_1\|_2.$$

We claim that in this case, we have

**Lemma 1.1.** Let $h$ be $M$-Lipschitz and $c$ have $L$-Lipschitz gradient. Then if $f(x) = h(c(x))$ and the model $f_x(y) = h(c(x) + \nabla c(x)^T(y - x))$, we have

$$|f(y) - f_x(y)| \leq \frac{LM}{2} \|x - y\|^2.$$  \hspace{1cm} (4)

We defer the proof of Lemma 1.1 to section 1.2. Thus the model-based iteration (2) (specialized to the composite case (3)) is a descent method whenever the stepsize $\alpha_k > 0$ is small enough that $\alpha_k \leq \frac{1}{\rho}$, where $\rho = L \|h\|_{\text{Lip}}$ is the product of the Lipschitz constants.

As an aside, this is part of a broader class of functions known as weakly convex functions that, while we do not particularly worry about them, is useful to at least define.

**Definition 1.1.** A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $\rho$-weakly convex if

$$f(x) + \frac{\rho}{2} \|x - x_0\|^2$$

is convex in $x$ for any $x_0 \in \mathbb{R}^n$.

In Def. 1.1, the choice of the point $x_0$ does not affect the definition in any way. Composite functions are $\rho$-weakly convex whenever $c$ has $L$-Lipschitz gradient, $h$ is $M$ Lipschitz, and $\rho \geq ML$. 

2
1.1 Examples of composite objectives

Here I collect a few examples, which may help; in the next section I give some basic pieces of analysis.

**Example 1** (Phase retrieval with $\ell_1$-loss): Let $A \in \mathbb{R}^{m \times n}$. The phase retrieval problem consists of attempting to recover a signal $x_\star$ from magnitude observations of the form

$$b_i = |\langle a_i, x_\star \rangle|^2, \quad i = 1, \ldots, m.$$  

A natural formulation is to consider the objective $f(x) = \frac{1}{m} \| (Ax)^2 - b \|_1$, where $\cdot^2$ denotes elementwise squaring. Then $f$ is evidently the composition of the convex function $h(z) = \frac{1}{m} \| z - b \|_1$ and the smooth function $c(x) = (Ax)^2$. Moreover, writing $c_i(x) = (a_i^T x)^2$, we have $\nabla c_i(x) = 2a_i a_i^T x$, and $\nabla^2 c_i(x) = 2a_i a_i^T$; as $c_i$ are each quadratic, we have

$$c_i(y) = c_i(x) + \nabla c_i(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 c_i(x) (y - x)$$

and, in particular,

$$|c_i(y) - b_i| = \left| c_i(x) + \nabla c_i(x)^T (y - x) - b_i + \frac{1}{2} (y - x)^T \nabla^2 c_i(x) (y - x) \right|$$

so that summing and using the triangle inequality, for the linearized model $f_x(y) = h(c(x) + \nabla c(x)^T (y - x))$ we have

$$|f_x(y) - f(y)| \leq \frac{1}{m} \sum_{i=1}^m (y - x)^T a_i a_i^T (y - x) = \frac{1}{m} (y - x)^T A^T A (y - x).$$

In particular, we can iterate the method

$$\Delta^k = \arg\min_{\Delta} \left\{ f_x(x^k + \Delta) + m^{-1} \Delta^T A^T A \Delta \right\}$$

$$x^{k+1} = x^k + \Delta^k,$$

which is guaranteed to be a descent method.

In the paper [1], we do a fairly deep investigation of Example 1, showing that it is very effective for solving phase retrieval problems.

**Example 2** (Matrix sensing): In matrix completion, we have observations of the form

$$b_i = \text{tr}(A_i^T X_\star), \quad i = 1, \ldots, m,$$

where $A_i \in \mathbb{R}^{n_1 \times n_2}$ and where $X_\star = x_\star y_\star^T$ for some $x_\star \in \mathbb{R}^{n_1}, y_\star \in \mathbb{R}^{n_2}$, so that $X_\star$ is assumed to be low rank. Then we can use the composite approaches above, where we let

$$c_i(x, y) = \text{tr}(A_i^T x y^T) = y^T A_i^T x, \quad \nabla c_i(x, y) = \begin{bmatrix} \nabla_x c_i(x, y) \\ \nabla_y c_i(x, y) \end{bmatrix} = \begin{bmatrix} A_i y \\ A_i^T x \end{bmatrix}.$$

We can write this in the composite objective form

$$f(x, y) = \| A(xy^T) - b \|_1 = \| c(x, y) - b \|_1,$$

where $A(X) = [\text{tr}(A_1^T X) \cdots \text{tr}(A_m^T X)]^T$ are the linear measurements we are taking.
Using the same linearization strategy as earlier, we can observe that the $c_i$ are quadratic functions, and moreover, their Hessians are exactly

$$\nabla^2 c_i(x, y) = \begin{bmatrix} 0 & A_i \\ A_i^T & 0 \end{bmatrix},$$

so that we have

$$f(x + \Delta x, y + \Delta y) = \sum_{i=1}^{m} |c_i(x + \Delta x, y + \Delta y) - b_i|$$

$$= \sum_{i=1}^{m} |c_i(x, y) - b_i + \nabla_x c_i(x, y)^T \Delta x + \nabla_y c_i(x, y)^T \Delta y + A_i^T \Delta y - A_i \Delta y|.$$ 

By the triangle inequality applied to $| \cdot |$, for the familiar model linearizing $c$

$$f_{(x, y)}(x', y') = h(c(x, y) + \nabla_x c(x, y)^T (x' - x) + \nabla_y c(x, y)^T (y' - y)),$$

which in this case specializes to

$$f_{(x, y)}(x + \Delta x, y + \Delta y) = \sum_{i=1}^{m} |c_i(x, y) - b_i + x^T A_i \Delta y + y^T A_i^T \Delta x|,$$

we have

$$f(x + \Delta x, y + \Delta y)|$$

$$\leq f_{(x, y)}(x + \Delta x, y + \Delta y) + \sum_{i=1}^{m} |\Delta_x A_i \Delta y|$$

$$= \|A(xy^T) - b + A(x \Delta_y^T) + A(\Delta_x y^T)\|_1 + \sum_{i=1}^{m} |\Delta_x A_i \Delta y|.$$ 

Note that, aside from the tedious algebra, we have basically only been applying inequality (4), because $| \cdot |$ is a 1-Lipschitz function, and each quadratic is quite smooth. We can specialize the above inequality even further in the case that the measurement matrices $A_i$ are low rank. In particular, suppose $A_i = u_i v_i^T$ for vectors $u_i \in \mathbb{R}^{n_1}, v_i \in \mathbb{R}^{n_2}$ with $\|u_i\|_2 = \|v_i\|_2 = 1$. Then for matrices $U = [u_1 \cdots u_m]^T$ and $V = [v_1 \cdots v_m]^T$, we have

$$A(x \Delta_y^T) = \text{diag}(U x) V \Delta_y \quad \text{and} \quad A(\Delta_x y^T) = \text{diag}(V y) U \Delta_x,$$

and moreover,

$$\Delta_x A_i \Delta y = \Delta_x^T u_i v_i^T \Delta y \leq \frac{1}{2} (u_i^T \Delta x)^2 + \frac{1}{2} (v_i^T \Delta y)^2,$$

where we used that $ab \leq \frac{1}{2} a^2 + \frac{1}{2} b^2$ for any $a, b \in \mathbb{R}$ (because $(a - b)^2 = a^2 + b^2 - 2ab \geq 0$), and in particular,

$$\sum_{i=1}^{m} |\Delta_x^T A_i \Delta y| \leq \frac{1}{2} \|U \Delta x\|_2^2 + \frac{1}{2} \|V \Delta y\|_2^2.$$

Summarizing, we have the inequality that

$$f(x + \Delta x, y + \Delta y) \leq \|A(xy^T) - b + \text{diag}(U x) V \Delta y + \text{diag}(V y) U \Delta x\|_1 + \frac{1}{2} \|U \Delta x\|_2^2 + \frac{1}{2} \|V \Delta y\|_2^2,$$

which gives a reasonably simple upper bound to compute."
1.2 Proof of Lemma 1.1

For each coordinate $i$ of $c(x) \in \mathbb{R}^m$ we have

$$c_i(y) = c_i(x) + \int_0^1 \nabla c_i(ty + (1-t)x)^T (y-x)dt$$

$$= c_i(x) + \nabla c_i(x)^T (y-x) + \int_0^1 (\nabla c_i(ty + (1-t)x) - \nabla c_i(x))^T (y-x)dt$$

by the fundamental theorem of calculus and that $\frac{\partial}{\partial t} c_i(ty + (1-t)x) = \nabla c_i(ty + (1-t)x)^T (y-x)$, and so

$$c(y) = c(x) + \nabla c(x)^T (y-x) + \int_0^1 (\nabla c(x + t(y-x)) - \nabla c(x))^T (y-x)dt.$$ 

Noting that

$$\|\nabla c(x + t(y-x)) - \nabla c(x)\|_2 \leq L \cdot t \|y-x\|_2,$$

the definition of the $\ell_2$-operator norm on matrices (i.e., that $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$ for any vector $x$) thus implies

$$\|c(y) - c(x) - \nabla c(x)^T (y-x)\|_2 \leq \int_0^1 t\|\nabla c(x + t(y-x)) - \nabla c(x)\|_2 |y-x|_2 dt$$

$$\leq L \int_0^1 t \|y-x\|_2(dt = \frac{L}{2} \|y-x\|^2_2.$$

In particular, we obtain that

$$|h(c(y)) - h(c(x) + \nabla c(x)^T(y-x))| \leq \|h\|_{\text{Lip}} \|c(y) - c(x) - \nabla c(x)^T(y-x)\|_2 \leq \frac{L \|h\|_{\text{Lip}}}{2} \|x-y\|^2_2$$

where $\|h\|_{\text{Lip}}$ denotes the Lipschitz constant of $h$.

References
