11 Bregman gradient methods

All the methods and convergence rates we have seen so far depend on the Euclidean structure we put on $\mathbb{R}^n$. For example, the smoothness and strong convexity assumptions we used are with respect to the Euclidean norm, and the obtained rates all involve a term of the form $\|x_0 - x^*\|_2$. In this lecture we will see that most of the results we have derived can be extended to work with so-called *Bregman divergences*.

11.1 Bregman divergence

Let $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a strictly convex[1] differentiable function, which is also lower semicontinuous[2]. The *Bregman divergence* associated to ϕ is the function:

$$D_\phi(x|y) = \phi(x) - [\phi(y) + \langle \nabla\phi(y), x - y\rangle]$$

defined for all $(x, y) \in \textbf{dom}\,\phi \times \textbf{int}\,\textbf{dom}\,\phi$. Convexity of ϕ tells us that $D_\phi(x|y) \geq 0$ for all x, y; and strict convexity tells us that $D_\phi(x|y) = 0 \implies x = y$.
Examples:

- If $\phi(x) = \|x\|_2^2/2$, then $D_\phi(x|y) = \|x\|_2^2/2 - \|y\|_2^2/2 - \langle y, x - y\rangle = \|x - y\|_2^2/2$ is the usual squared Euclidean norm.
- If $\phi(x) = \sum_{i=1}^n x_i \log x_i$ defined on $\mathbb{R}_+^n$, then

$$D_\phi(x|y) = \sum_{i=1}^n x_i \log(x_i/y_i) + y_i - x_i$$

is the so-called *Kullback-Leibler (KL) divergence*, defined for all $x \geq 0$ and $y > 0$.

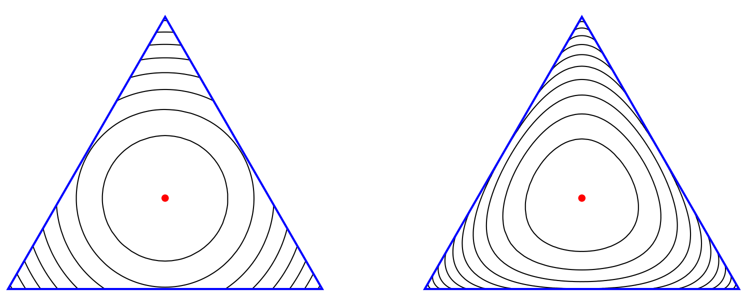

Figure 1: Contour plots of $\|x - p\|_2^2/2$ vs. $D_{KL}(x|p)$, where $p = (1/3, 1/3, 1/3)$, on the unit simplex $\{x \in \mathbb{R}^3 : x \geq 0 \text{ and } x_1 + x_2 + x_3 = 1\}$.

[1] A strictly convex function is one that satisfies $\phi(\lambda x + (1 - \lambda)y) < \lambda\phi(x) + (1 - \lambda)\phi(y)$ for all x, y and $\lambda \in (0, 1)$.
[2] Recall that ϕ is lower semicontinuous iff all its sublevel sets are closed.

EXERCISE: Show, using strict convexity of ϕ, that the balls $\{x \in \mathbf{dom}(\phi) : D_\phi(x|y) \leq r\}$ for any $y \in \mathbf{int\,dom}\,\phi$ and any $r \geq 0$ are all bounded. [Hint: you can use the fact that if C is an unbounded closed convex set, then there is a direction v such that $x + tv \in C$ for all $x \in C$ and $t \geq 0$.]

We will need the following identity, which is straightforward to verify. This identity generalizes the following "completion of squares" identity, which we have used repeatedly in previous convergence proofs:

$$\|c - b\|_2^2 - 2\langle c - b, a - b\rangle = \|c - a\|_2^2 - \|b - a\|_2^2.$$

Proposition 11.1. *For any a, b, c we have*

$$D_\phi(c|b) - \langle \nabla\phi(a) - \nabla\phi(b), c - b\rangle = D_\phi(c|a) - D_\phi(b|a). \tag{1}$$

The following figure gives a simple graphical intepretation of this equality.

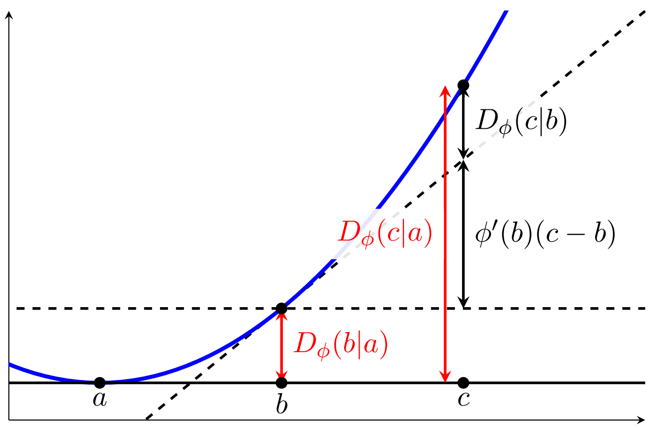

Figure 2: Illustration of the equality (1) for a univariate function ϕ, where $\phi'(a) = 0$.

When $c = a$ the identity (1) tells us that

$$\langle \nabla\phi(a) - \nabla\phi(b), a - b\rangle = D_\phi(a|b) + D_\phi(b|a). \tag{2}$$

11.2 Bregman gradient method

Consider the problem of minimizing $f(x)$ over $x \in \mathbb{R}^n$, where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. We have seen that the iterates of the gradient method can be expressed in the following way:

$$x_{k+1} = x_k - t_k\nabla f(x_k) = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{1}{2t_k}\|x - x_k\|_2^2\right\}.$$

The *Bregman gradient method* (a.k.a. *mirror descent*) corresponds to replacing the term $\|x - x_k\|_2^2/2$ by a general Bregman divergence generated by ϕ, i.e., it takes the form

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{1}{t_k}D_\phi(x|x_k)\right\}. \tag{3}$$

Remarks:

- The optimality condition for the minimization expression above tells us that we must have

$$\nabla f(x_k) = -\frac{1}{t_k}(\nabla\phi(x_{k+1}) - \nabla\phi(x_k)). \tag{4}$$

  Compare this with the identity $\nabla f(x_k) = -\frac{1}{t}(x_{k+1} - x_k)$ we used when analyzing the gradient method.

- Equation (4) can also be rewritten as

$$x_{k+1} = (\nabla\phi)^{-1}(\nabla\phi(x_k) - t_k\nabla f(x_k)). \tag{5}$$

  The function $\nabla\phi$ maps vectors in $\mathbb{R}^n$ to vectors in the dual space. The operation $\nabla\phi(x_k) - t_k\nabla f(x_k)$ is carried out in the dual space of $\mathbb{R}^n$, and the operation $(\nabla\phi)^{-1}$ is used to map the iterates back to the primal space $\mathbb{R}^n$. In the form (5), these iterations are known as *mirror descent method.*

Example 1. *Consider the problem of minimizing $f(x)$ on $\mathbb{R}^n_+$. If we choose $D_\phi = D_{KL}$ the KL-divergence, then the iterates are defined by $x_{k+1} = \text{argmin}_{x\geq 0}\{t_k \langle\nabla f(x_k), x - x_k\rangle + D_{KL}(x|x_k)\}$ which can be shown to be equal to*

$$x_{k+1} = x_k \bullet \exp(-t_k\nabla f(x_k))$$

where $\bullet$ denotes componentwise multiplication, and exp *here is the componentwise exponential function. This iteration is known as* exponentiated gradient descent*.*

The analysis of the gradient method can be adapted to the case of the Bregman gradient method provided we use the following assumptions on f.

Definition 11.1 (Relative smoothness, and relative strong convexity)**.** Let $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a strictly convex function. We say that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *L-smooth relative to ϕ* if $L\phi - f$ is convex. We say that f is *m-strongly convex relative to ϕ* if $f - m\phi$ is convex.

Remark 1. *When $\phi(x) = \|x\|_2^2/2$, then we recover the notions of L-smoothness and m-strong convexity with respect to the Euclidean norm.*

Equipped with the definitions above, we can prove the following.

Theorem 11.1. *If f is convex and L-smooth relative to ϕ, then the iterates of the Bregman gradient method (3) with constant step size $t_k = t \in (0, 1/L]$ satisfy for all $k \geq 1$.*

$$f(x_k) - f^* \leq \frac{1}{kt}D_\phi(x^*|x_0). \tag{6}$$

If, in addition, f is m-strongly relative to ϕ, then we have for all $k \geq 1$

$$D_\phi(x^*|x_k) \leq (1 - mt)^k\, D_\phi(x^*|x_0). \tag{7}$$

Proof. We start by proving (6). The proof follows the same line as the proofs we have seen before. The assumption that $L\phi - f$ is convex tells us that $D_{L\phi-f} \geq 0$, which corresponds to

$$f(x_{k+1}) \leq f(x_k) + \langle\nabla f(x_k), x_{k+1} - x_k\rangle + LD_\phi(x_{k+1}|x_k). \tag{8}$$

(Compare with the descent lemma.) We substract $f(u)$ from each side of (8) and use convexity of f to get

$$\begin{aligned} f(x_{k+1}) - f(u) &\leq f(x_k) - f(u) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + L D_\phi(x_{k+1}|x_k) \\ &\leq \langle \nabla f(x_k), x_k - u \rangle + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + L D_\phi(x_{k+1}|x_k) \\ &= \langle \nabla f(x_k), x_{k+1} - u \rangle + L D_\phi(x_{k+1}|x_k). \end{aligned} \tag{9}$$

Using the expression (4) for $\nabla f(x_k)$ we get

$$f(x_{k+1}) - f(u) \leq -(1/t) \langle \nabla \phi(x_{k+1}) - \nabla \phi(x_k), x_{k+1} - u \rangle + L D_\phi(x_{k+1}|x_k). \tag{10}$$

The three-point identity (1) with $a = x_k, b = x_{k+1}, c = u$ tells us that

$$\langle \nabla \phi(x_{k+1}) - \nabla \phi(x_k), x_{k+1} - u \rangle = D_\phi(u|x_{k+1}) - D_\phi(u|x_k) + D_\phi(x_{k+1}|x_k).$$

Plugging this in (10) and using the fact that $t \leq 1/L$ we get

$$f(x_{k+1}) - f(u) \leq (-1/t)(D_\phi(u|x_{k+1}) - D_\phi(u|x_k)). \tag{11}$$

Taking $u = x_k$ tells us that we are dealing with a descent method, i.e., $f(x_{k+1}) \leq f(x_k)$. Taking $u = x^*$, and summing the inequalities from $k = 0$ to $k = K - 1$ gives us

$$K(f(x_K) - f(x^*)) \leq \sum_{k=0}^{K-1} f(x_{k+1}) - f(x^*) \leq (-1/t)(D_\phi(x^*|x_K) - D_\phi(x^*|x^0)) \leq \frac{1}{t} D_\phi(x^*|x^0).$$

Dividing by K gives us the desired inequality (6).

The proof of (7) is very similar. The difference is that in (9), we write the *equality* $f(x_k) - f(u) = \langle \nabla f(x_k), x - u \rangle - D_f(u|x_k)$, and then, since $f - m\phi$ is convex, we have $D_{f-m\phi} = D_f - mD_\phi \geq 0$, and so we can write $D_f(u|x_k) \geq mD_\phi(u|x_k)$. The inequality (11) then becomes

$$\begin{aligned} f(x_{k+1}) - f(u) &\leq (-1/t)(D_\phi(u|x_{k+1}) - D_\phi(u|x_k)) - mD_\phi(u|x_k) \\ &= -(1/t)D_\phi(u|x_{k+1}) + (1/t - m)D_\phi(u|x_k). \end{aligned}$$

Taking $u = x^*$, and using the fact that $0 \leq f(x_{k+1}) - f(x^*)$, we get

$$D_\phi(x^*|x_{k+1}) \leq (1 - mt)D_\phi(x^*|x_k)$$

as desired. □

Remark 2. *The assumption $L\phi - f$ convex was introduced in [BBT17] as the Lipschitz-like/Convexity condition, also known as* relative smoothness *in [LFN18].*

References

[BBT17] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. 4

[LFN18] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. 4

[Teb18] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.