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 convergence rates we have derived 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 $\phi : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a smooth, strictly[1] convex 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 $(x, y) \in \mathbf{dom}\,\phi \times \mathbf{int\,dom}\,\phi$. 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 $D_\phi(x|y) = \|x\|_2^2/2 - \|y\|_2/2 - \langle y, x - y\rangle = \|x - y\|_2^2/2$ 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 $y > 0$.

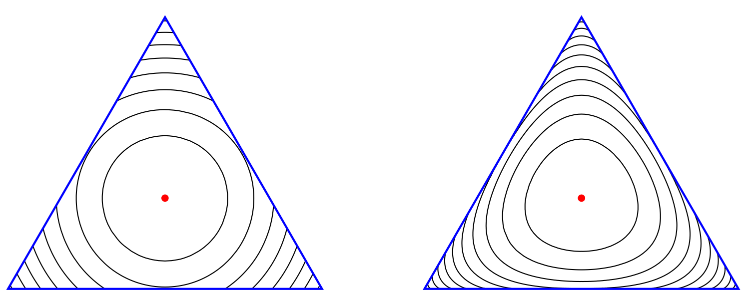

Figure 1: Contour plots of $\|x - p\|_2^2/2$ vs. $D_{KL}(x|p)$, where $p = (1/3, 1/3, 1/3)$, on the unit simplex $\{x \in \mathbb{R}^3 : x \geq 0 \text{ and } x_1 + x_2 + x_3 = 1\}$.

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

[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.

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.

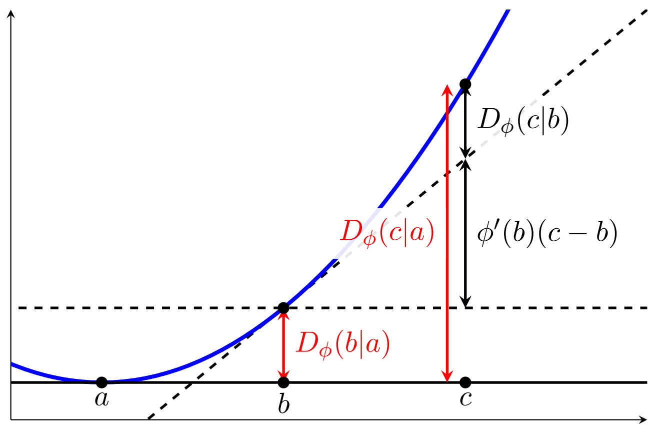

Figure 2: Illustration of the equality (1) for a univariate function ϕ, where $\phi'(a) = 0$.

11.2 Bregman proximal operator

We define the Bregman proximal operator for a function f to be:

$$\mathbf{prox}_f^\phi(y) = \operatorname*{argmin}_{x\in\mathbb{R}^n} \{f(x) + D_\phi(x|y)\}.$$

When $\phi(x) = \|x\|_2^2/2$, this is the proximal operator we saw in the previous lecture. Under mild conditions (e.g., $\mathbf{int\,dom}\, f \subset \mathbf{int\,dom}\,\phi \neq \emptyset$), we have:

$$x = \mathbf{prox}_f^\phi(y) \iff 0 \in \partial f(x) + \nabla\phi(x) - \nabla\phi(y). \tag{2}$$

EXERCISE: Show that if f is lower semicontinuous and bounded below, then $\mathbf{prox}_f^\phi(y)$ is well-defined.

Properties of the proximal operator We saw that the usual proximal operator is a (firmly) nonexpansive operator. Here, the generalized prox operator satisfies a certain nonexpansive property, but only wrt minimizers.

Proposition 11.2. *Let $x = \mathbf{prox}_f^\phi(y)$. Then for any u, we have*

$$f(u) + D_\phi(u|y) \geq f(x) + D_\phi(x|y) + D_\phi(u|x).$$

Note that the inequality would be trivial if we did not have the last term $D_\phi(u|x)$.

Proof. From (2), we know that $x = \mathbf{prox}_f^\phi(y)$ if, and only if, $\nabla\phi(y) - \nabla\phi(x) \in \partial f(x)$. Thus this means that for any u we have:

$$f(u) \geq f(x) + \langle \nabla\phi(y) - \nabla\phi(x), u - x \rangle$$

By the three-point identity (1) with $a = x, b = u, c = y$, we have $\langle \nabla\phi(y) - \nabla\phi(x), u - x \rangle = D_\phi(x|y) + D_\phi(u|x) - D_\phi(u|y)$, which gives the desired result. □

11.3 Bregman proximal gradient algorithm

Consider the problem of minimizing $F(x) = f(x) + h(x)$ over $x \in \mathbb{R}^n$, where $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is smooth, and $h : \mathbb{R}^n \to \bar{\mathbb{R}}$ has a *simple prox*. The Bregman proximal gradient method, takes the following form:

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

where $t > 0$ is the step size. If $D_\phi(x|x_k) = (1/2)\|x - x_k\|_2^2$, then this precisely the iteration $x_{k+1} = \mathbf{prox}_{th}(x_k - t\nabla f(x_k))$ that we saw in the last lecture. When $h(x) = 0$, this method is *not* the usual gradient method.

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

Convergence We have previously studied the convergence of the (standard) proximal gradient method under the assumption that f is L-smooth. Recall that L-smoothness is equivalent to having $L\|x\|_2^2 - f(x)$ convex. It is thus natural to study the convergence of the Bregman gradient method under the assumption that $L\phi - f$ is convex. In fact, we can show that if $L\phi - f$ is convex, then the iterates of the Bregman proximal gradient method (3) with step size $t = 1/L$ satisfy

$$F(x_k) - F^* \leq \frac{L D_\phi(x^*|x_0)}{k}. \tag{4}$$

Note that this is precisely the same convergence result we obtained before, where the term $\|x^* - x_0\|_2^2$ is now replaced by $D_\phi(x^*|x_0)$.

Proof: The proof follows more or less 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

$$f(x^+) \leq f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + L D_\phi(x^+|x).$$

We substract $f(u)$ from each side to get

$$\begin{aligned} f(x^+) - f(u) &\leq f(x) - f(u) + \left\langle \nabla f(x), x^+ - x \right\rangle + L D_\phi(x^+|x) \\ &\overset{(*)}{\leq} \langle \nabla f(x), x - u \rangle + \left\langle \nabla f(x), x^+ - x \right\rangle + L D_\phi(x^+|x) \\ &= \left\langle \nabla f(x), x^+ - u \right\rangle + L D_\phi(x^+|x), \end{aligned} \tag{5}$$

where in (*) we have used convexity of f. So far we haven't used that x^+ is computed from (3). Using Prop. 11.2, with the function $z \mapsto t\langle \nabla f(x), z - x\rangle + th(z)$, we get that

$$t\left\langle \nabla f(x), x^+ - u\right\rangle + t(h(x^+) - h(u)) \leq D_\phi(u|x) - D_\phi(u|x^+) - D_\phi(x^+|x).$$

Using $t = 1/L$, and plugging in (5) we finally reach

$$t(F(x^+) - F(u)) \leq D_\phi(u|x) - D_\phi(u|x^+)$$

as desired. The rest of the proof is as usual: letting $u = x$ we see that $F(x^+) \leq F(x)$. Then we let $u = x^*$, sum the inequalities from 0 to $k-1$ to reach the desired inequality (4).

Remark 1. *Another approach to proving convergence is to remark that since $\phi - tf$ is convex, the iteration* (3) *can be written as:*

$$x_{k+1} = \mathbf{prox}_{t(f+h)}^{\phi - tf}(x_k).$$

Using Prop 11.2 we get

$$t(F(x^+) - F(u)) \leq D_{\phi - tf}(u|x) - D_{\phi - tf}(u|x^+).$$

Proceeding as usual, we get $F(x_k) - F^* \leq LD_{\phi - tf}(x^*|x_0)/k \leq LD_\phi(x^*|x^0)/k$ as desired.

Strongly convex case: For the standard proximal gradient, we saw that if f is m-strongly convex (i.e., $f - m\|\cdot\|_2^2/2$ is convex), then convergence is linear with a rate of $1 - m/L$. The same can be proved here. Indeed, we can show that if $f - m\phi$ is convex, then the iterates (3) with $t = 1/L$ satisfy

$$D_\phi(x^*|x_k) \leq \left(1 - \frac{m}{L}\right)^k D_\phi(x^*|x_0).$$

The proof is a simple modification to the proof we just saw: in step (*) of (5), we write the *equality* $f(x) - f(u) = \langle \nabla f(x), x - u\rangle - D_f(u|x)$, and then, since $f - m\phi$ is convex, we upper bound the second term using $D_f(u|x) \geq mD_\phi(u|x)$. Continuing with the same steps as before, we eventually get

$$t(F(x^+) - F(u)) \leq (1 - mt)D_\phi(u|x) - D_\phi(u|x^+).$$

With $u = x^*$, the left-hand side is nonnegative, and so we get $D_\phi(u|x^+) \leq (1 - mt)D_\phi(u|x)$ which gives us the linear convergence rate.

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.