12 Mirror descent (non-Euclidean gradient method)

In this lecture we come back to the problem of minimizing a nonsmooth convex function on a closed convex set C:

$$\min_{x \in C} f(x).$$

Projected subgradient iterations work as follows:

$$x_{k+1} = P_C(x_k - t_k g_k) \tag{1}$$

where $g_k \in \partial f(x_k)$ and P_C is the Euclidean projection on C. Analysis of this method with suitable step size (e.g., Polyak step size) guarantees that after k iterations we have

$$f_{\text{best},k} - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{k+1}} \tag{2}$$

where $f_{\text{best},k} = \min\{f(x_0), \ldots, f(x_k)\}$ and $G = \max\{\|g_0\|_2, \ldots, \|g_k\|_2\}$.

Dependence on Euclidean inner product The subgradient method (1) depends on the Euclidean structure we put on $\mathbb{R}^n$. This is apparent from the use of the Euclidean projection P_C, and from the identification of the subgradient as an element of $\mathbb{R}^n$ (rather than as an element of the dual space). This dependence on the Euclidean structure is reflected in the analysis (2) where G is the Lipschitz constant of f wrt Euclidean norm.

Mirror descent We consider in this lecture a non-Euclidean version of subgradient descent, known as *mirror descent*. We fix a smooth convex function ϕ defined on the convex set $C \subset \mathbb{R}^n$. We assume ϕ is 1-strongly convex wrt some norm $\|\cdot\|$ on $\mathbb{R}^n$, i.e.,

$$\phi(x) \geq \phi(y) + \nabla\phi(y)^T(x-y) + \frac{1}{2}\|x-y\|^2 \qquad \forall x, y \in \text{dom}(\phi).$$

The *Bregman divergence* associated to ϕ is

$$D_\phi(x\|y) = \phi(x) - \phi(y) - \nabla\phi(y)^T(x-y).$$

Note that $D_\phi(x\|y) \geq \|x-y\|^2/2$ by strong convexity of ϕ. The mirror descent algorithm is defined by the iterates:

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ t_k g_k^T(x - x_k) + D_\phi(x\|x_k) \right\}. \tag{3}$$

Remark. *One can easily check that with $\phi(x) = \|x\|_2^2/2$ we recover the subgradient method (1). Indeed iterate (1) can be equivalently written as $x_{k+1} = \operatorname{argmin}_{x \in C} \left\{ t_k g_k^T(x - x_k) + \|x - x_k\|_2^2/2 \right\}$.*

Analysis of mirror descent With a suitable choice of step size, one can show that the sequence (x_k) produced by (3) satisfies:

$$f_{\text{best},k} - f^* \leq \frac{G\sqrt{D_\phi(x^*\|x_0)}}{\sqrt{k+1}}$$

where G is the Lipschitz constant of f wrt $\|\cdot\|$. We prove this now.

Bregman projection Let C be a closed convex set and let $x_0 \notin C$. We know that if $\hat{x}$ is the Euclidean projection of x_0 on C, then for any $x \in C$ we have the following inequality

$$\|x - x_0\|_2^2 \;\; \geq \;\; \|\hat{x} - x_0\|_2^2 + \|x - \hat{x}\|_2^2. \tag{4}$$

This inequality is another way of saying that the angle at $\hat{x}$ formed by the vectors $x_0 - \hat{x}$ and $x - \hat{x}$ is obtuse (see Figure 1).

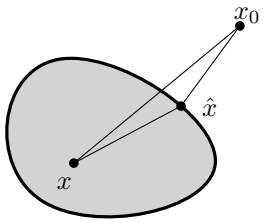

Figure 1: If $\hat{x}$ is the Euclidean projection of x_0 on the convex set C, then the following inequality holds for all $x \in C$: $\|x - x_0\|_2^2 \geq \|\hat{x} - x_0\|_2^2 + \|x - \hat{x}\|_2^2$ (obtuse angle at $\hat{x}$). The previous inequality generalizes to projections defined using general Bregman divergences.

Inequality (4) generalizes to projections defined using the Bregman divergence. More precisely, if we define

$$\hat{x} = \operatorname*{argmin}_{x \in C} D_\phi(x \| x_0),$$

then one can show that the following inequality holds, for any $x \in C$:

$$D_\phi(x \| x_0) \;\; \geq \;\; D_\phi(\hat{x} \| x_0) + D_\phi(x \| \hat{x}). \tag{5}$$

This is proved in the following lemma, in the more general functional setting (take g to be the indicator function of C to recover (5)).

Lemma 1. *Consider the optimization problem* $\min_{x \in \mathbb{R}^n} \{g(x) + D_\phi(x \| x_0)\}$. *Point* $\hat{x}$ *is optimal iff*

$$g(x) + D_\phi(x \| x_0) \;\; \geq \;\; g(\hat{x}) + D_\phi(\hat{x} \| x_0) + D_\phi(x \| \hat{x})$$

for all $x \in \mathrm{dom}(g)$.

Proof. This lemma would be trivial if we didn't have the last term $D_\phi(x \| \hat{x})$. The whole point of this lemma is this last term. Using the definition of D_ϕ note that $\hat{x} \in \operatorname{argmin} \left\{ g(x) + \phi(x) - \nabla\phi(x_0)^T x \right\}$. This means that $0 \in \partial g(\hat{x}) + \nabla\phi(\hat{x}) - \nabla\phi(x_0)$, or equivalently

$$-\nabla\phi(\hat{x}) \in \partial g(\hat{x}) - \nabla\phi(x_0) = \partial(g - \nabla\phi(x_0)^T \cdot)(\hat{x}). \tag{6}$$

This implies, by definition of subgradient, that we have for any x:

$$g(x) - \nabla\phi(x_0)^T x \;\geq\; g(\hat{x}) - \nabla\phi(x_0)^T \hat{x} - \nabla\phi(\hat{x})^T (x - \hat{x}). \tag{7}$$

Now we finish the proof. We have, for any x:

$$\begin{aligned} g(x) + D_\phi(x \| x_0) - g(\hat{x}) - D_\phi(\hat{x} \| x_0) &= (g(x) - \nabla\phi(x_0)^T x) - (g(\hat{x}) - \nabla\phi(x_0)^T \hat{x}) + \phi(x) - \phi(\hat{x}) \\ &\overset{\text{by (7)}}{\geq} -\nabla\phi(\hat{x})^T (x - \hat{x}) + \phi(x) - \phi(\hat{x}) = D_\phi(x \| \hat{x}). \end{aligned}$$

□

We are now ready to proceed with the analysis of mirror descent iterations. Our goal will be to bound $D_\phi(x^*\|x_{k+1})$ in terms of $D_\phi(x^*\|x_k)$. We apply Lemma 1 to (3) where $g(x) = I_C(x) + t\nabla f(x_k)^T(x - x_k)$, with I_C being the indicator function of C. Since $x^* \in C$ we have

$$t\nabla f(x_k)^T(x^* - x_k) + D_\phi(x^*\|x_k) \;\geq\; t\nabla f(x_k)^T(x_{k+1} - x_k) + D_\phi(x_{k+1}\|x_k) + D_\phi(x^*\|x_{k+1}).$$

Rearranging, this tells us

$$\begin{aligned}
D_\phi(x^*\|x_{k+1}) &\leq D_\phi(x^*\|x_k) - D_\phi(x_{k+1}\|x_k) + t\nabla f(x_k)^T(x_k - x_{k+1}) + t\nabla f(x_k)^T(x^* - x_k) \\
&\overset{(a)}{\leq} D_\phi(x^*\|x_k) - D_\phi(x_{k+1}\|x_k) + \|t\nabla f(x_k)\|_*\|x_k - x_{k+1}\| + t\nabla f(x_k)^T(x^* - x_k) \\
&\overset{(b)}{\leq} D_\phi(x^*\|x_k) - D_\phi(x_{k+1}\|x_k) + \frac{1}{2}\|t\nabla f(x_k)\|_*^2 + \frac{1}{2}\|x_k - x_{k+1}\|^2 + t\nabla f(x_k)^T(x^* - x_k) \\
&\overset{(c)}{\leq} D_\phi(x^*\|x_k) + \frac{1}{2}\|t\nabla f(x_k)\|_*^2 + t(f^* - f(x_k)).
\end{aligned}$$

where in (a) we used the (generalized) Cauchy-Schwarz inequality, in (b) we used the arithmetic-geometric mean inequality, in (c) we used that $D_\phi(a\|b) \geq \frac{1}{2}\|a - b\|^2$ (strong convexity of ϕ) and convexity of f. We now finish the proof like we did for the subgradient method. We apply the inequality recursively to get at the end

$$D_\phi(x^*\|x_{k+1}) \;\leq\; D_\phi(x^*\|x_0) + \sum_{i=0}^{k} t_i^2\|\nabla f(x_i)\|_*^2 + \sum_{i=0}^{k} t(f^* - f(x_i))$$

which, after rearranging gives

$$\sum_{i=0}^{k} t_i(f(x_i) - f^*) \leq D_\phi(x^*\|x_0) + \sum_{i=0}^{k} t_i^2\|\nabla f(x_i)\|_*^2.$$

Define $f_{\text{best},k} = \min\{f(x_0), \ldots, f(x_k)\}$ to get

$$f_{\text{best},k} - f^* \leq \frac{D_\phi(x^*\|x_0)}{\sum_{i=0}^{k} t_i} + \frac{\sum_{i=0}^{k} t_i^2\|\nabla f(x_i)\|_*^2}{\sum_{i=0}^{k} t_i}.$$

The rest of the proof is like with subgradient method.