Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Victor Ye Dong
Email: yd2470@columbia.edu

SCRIBE
19

Analysis of Mirror Descent

1 Basic Notions

Mirror descent is a generalization of GD. The method is adapted to a norm $\|\cdot\|$ on $\mathbb{R}^d$ that is appropriate to the geometry of the problem. The problem we want to solve is $\min_{x\in K} f(x)$, where f is a convex function and K is a convex set as usual. The difference from standard GD is that we now impose bounds on K and the gradients of f using a (potentially non-Euclidean) norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$ (definition to follow). The dual norm is defined on the dual vector space (i.e. the vector space of all linear functionals on the original vector space). The finite dimension Euclidean space R^d is self-dual; i.e. the dual is also $\mathbb{R}^d$, so for the purpose of this lecture, we won't need to worry about vector space duality.

Definition 1 (Dual Norm)**.**

$$\|x\|_* = \sup_{\|v\|\leq 1} |x \cdot v| = \sup_{v\neq 0} \frac{|x \cdot v|}{\|v\|} \tag{1.1}$$

Example 1: Dual norm for ℓ_2 is ℓ_2 :

$$\|x\|_2 = \sup_{\|v\|_2\leq 1} |x \cdot v| \tag{1.2}$$

$\Rightarrow$

$$v^* = \frac{x}{\|x\|_2} \Rightarrow |x \cdot v| = \frac{\|x\|_2^2}{\|x\|_2} = \|x\|_2 \tag{1.3}$$

Example 2 : Dual norm of ℓ_p is ℓ_q where $\frac{1}{p} + \frac{1}{q} = 1$. Interesting dual pairs are $(2, 2)$ and $(1, \infty)$, which will be our main focus below.

Fact : For $x \in v, y \in v^*, x, y \in \mathbb{R}^d$,

$$|x \cdot y| \leq \|x\|\|y\|_* \quad \text{generalized Cauchy-Schwartz} \tag{1.4}$$

Definition 2 (Mahalanobis Distance(Generalization of ℓ_2 norm))**.** If $A \succ 0, \|x\|_A = \sqrt{x^T Ax} = \|A^{\frac{1}{2}}x\|_2$

Fact :

$$\|x\|_A, * = \|x\|_{A^{-1}} = \sqrt{x^T A^{-1} x} \tag{1.5}$$

2 Mirror Map and Bregman divergence

In order to define mirror descent, we need two notions: mirror map and Bregman divergence.

Definition 3. For an open set $U \subseteq \mathbb{R}^d$, the function $\phi : U \to \mathbb{R}$ is mirror map if

a. ϕ : 1 - strongly convex in $\|\cdot\|$ norm.

$$\phi(x') \geq \phi(x) + \nabla\phi(x)(x' - x) + \frac{1}{2}\|x - x'\|^2 \tag{2.1}$$

b. $\nabla\phi$: $U \to \mathbb{R}^d$ surjective (or in other words, the inverse function $[\nabla f]^{-1} : \mathbb{R}^d \to U$ is well-defined).

c. $\|\nabla\phi(x)\| \to \infty$ when $x \to \partial U$

We mainly discuss two cases in the following analysis,

1) For the ℓ_2 norm case, consider $\phi(x) = \frac{1}{2}\|x\|^2$ defined on $U = \mathbb{R}^d$. Then we have $\nabla\phi(x) = x, \nabla^2\phi(x) = I$, so ϕ is indeed a valid mirror map for the ℓ_2 norm.

2) For the ℓ_1 norm case, let $U = \{x \in \mathbb{R}^d | x_i > 0 \forall i\}$, and define $\phi(x) = \sum_i x_i \log(x_i)$, the *negative entropy function.* Then $\nabla\phi(x) = \log(x_i) + 1$. It is easy to check that $\nabla\phi$ is surjective: given a vector $v \in \mathbb{R}^d$, if we define

$$x_i = e^{v_i - 1} \tag{2.2}$$

then $\nabla\phi(x) = v$. The strong convexity of $\phi(x)$ follows from Pinsker's inequality.

Definition 4 (Bregman Divergence)**.** For a convex function $\phi : U \to \mathbb{R}$, the Bregman divergence between a pair of points $x, x' \in U$ is defined as

$$B_\phi(x', x) = \phi(x') - (\phi(x) + \nabla\phi(x) \cdot (x' - x)) \tag{2.3}$$

The Bregman divergence $B_\phi(x'x)$ measures how much $\phi(x')$ diverges from the linear approximation to ϕ at x. Note that the Bregman divergence is *not* symmetric between x and x', so the order of the arguments matters.

For the ℓ_2 norm, using the mirror map $\phi(x) = \frac{1}{2}\|x\|_2^2$, we have

$$\begin{aligned} B_\phi(x', x) &= \frac{1}{2}\|x'\|^2 - (\frac{1}{2}\|x\|^2 + x(x' - x)) \\ &= \frac{1}{2}\|x - x'\|^2 \end{aligned} \tag{2.4}$$

For the ℓ_1 case, using the mirror map $\phi(x) = \sum_i x_i \log(x_i)$, we have

$$\begin{aligned} B_\phi(x', x) &= \sum_i x_i' \log(x_i') - (\sum_i x_i \log(x_i) + (\sum_i \log(x_i) + 1)(x_i' - x_i)) \\ &= \sum_i x_i' \log\left(\frac{x_i'}{x_i}\right) - \sum_i (x_i' - x_i) \end{aligned} \tag{2.5}$$

Note that in the special case when x and x' are distributions over the d coordinates (i.e. $\sum_i x_i = \sum_i x_i' = 1$) the Bregman divergence becomes exactly the Kullback-Leibler (KL) divergence, or

relative entropy, between x' and x, since the term $\sum_i (x'_i - x_i)$ vanishes.

3 Mirror Descent

Mirror descent is an elegant method to exploit the geometry of the problem when the geometry is best described by some (not necessarily Euclidea) norm $\|\cdot\|$. Suppose now that we are given constants $D > 0$ and $L > 0$ such that $\|x\| \leq D$ and $\|\nabla f(x)\|_* \leq L$ for all $x \in K$ (here, we are interpreting $\nabla f(x)$ as a member of the dual to $\mathbb{R}^d$). The algorithm is given below. In words, the algorithm uses a mirror map ϕ adapted to the norm $\|\cdot\|$ to map the current iterate x_t into the dual space as $\nabla\phi(x_t)$. This is the same space that the gradient $\nabla f(x_t)$ lives in, so we can take a gradient descent step in the dual step as $\nabla\phi(x_t) - \eta\nabla f(x_t)$. To map this back to the primal space where the x's live, we apply the inverse mirror mapping, i.e. we compute $y_{t+1} = [\nabla\phi]^{-1}(\nabla\phi(x_t) - \eta\nabla f(x_t))$. This point may not lie in K, so to bring the new iterate back into K, we perform a projection: here the appropriate projection is the *Bregman* projection, i.e. we compute $x_{t+1} = \arg\min_{x\in K} B_\phi(x, y_{t+1})$.‘

Algorithm 1: Mirror Descent

Start with an arbitrary $x_0 \in K$.

for $t = 0, 1, 2, \ldots, T$ **do**

$y_{t+1} = [\nabla\phi]^{-1}(\nabla\phi(x) - \eta\nabla f(x_t))$

$x_{t+1} = \arg\min_{x\in K} B_\phi(x, y_{t+1})$

end

Output $\frac{1}{T}\sum_{i=1}^{T-1} x_i$

4 Instantiating the algorithm for various special cases

For the ℓ_2 norm case, it is easy to check that mirror descent reduces exactly to standard projected gradient descent.

For ℓ_1 norm case, we have $\phi(x) = \sum_i x_i \log(x_i)$, for which

$$\nabla\phi(x) = \langle \log(x_i) + 1\rangle_{i=1}^d \text{ and } [\nabla\phi]^{-1}(v) = \langle e^{v_i - 1}\rangle_{i=1}^d$$

Thus, we have

$$\begin{aligned} y_{t+1,i} &= \exp(\log(x_{t,i}) + 1 - \eta\nabla f(x_t)_i - 1) \\ &= \exp(\log(x_{t,i}) - \eta\nabla f(x_t)_i) \\ &= x_{t,i} \cdot exp(-\eta\nabla f(x_t)_i). \end{aligned} \tag{4.1}$$

Then, to compute

$$x_{t+1} = \arg\min_{x\in K} B_\phi(x, y_{t+1}) \tag{4.2}$$

one can check (via Lagrange multipliers) that the Bregman projection amounts to re-normalizing y_{t+1} so that the coordinates sum to 1, i.e. $x_{t+1} = \frac{y_{t+1}}{\|y_{t+1}\|_1}$.