CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 23**

Online convex optimization: mirror descent

Lecturer: Sasha Rakhlin *Scribe: Lei Shi*

1 Bregmen Divergence (cont.)

Bregman divergence with respect to $\mathbb{R}$:

$$D_R(x, y) = R(x) - R(y) - \nabla R(y)(x - y) \tag{1}$$

(cont.) Properties of Bregman divergence

- 5, Define Legendre dual
$$R^*(u) = \sup_v [u \cdot v - R(v)] \tag{2}$$
examples: $R(x) = \frac{1}{2}||x||_p^2 \leftrightarrow R^*(x) = \frac{1}{2}||x||q^2$, where $\frac{1}{p} + \frac{1}{p} = 1$
- 6,$\nabla R^* = (\nabla R)^{-1}$
- 7, $D_R(u, v) = D_{R^*}(\nabla R(x), \nabla R(u))$
- 8, $D_{R+f}(x, y) = D_R(x, y)$, if $f(x)$ is linear
- 9, $\nabla_x D_R(x, y) = \nabla R(x) - \nabla R(y)$
- 10, If y minimize R ($\nabla R(y) = 0$) then $D_R(x, y) = R(x) - R(y)$

2 Recap: online convex optimization

For t=1:T

- Player choose $x_t \in K$ (convex)
- Adversary choose $l_t(\cdot)$ (convex)

Goal: minimize regret

$$R_T = \sum_{t=1}^{T} l_t(x_t) - \min_{u \in K} \sum_{t=1}^{T} l_t(u) \tag{3}$$

Consider the following family of algorithms:

$$x_{t+1} = \arg\min_{x \in K} \eta \sum_{s=1}^{t} l_t(x) + R(x) \tag{4}$$

for some $R(\cdot)$ convex.

Define $\Phi_0(x) := R(x), \Phi_t(x) := \Phi_{t-1}(x) + \eta l_t(x)$

Lemma 2.1. Suppose $K = \mathbb{R}^n$, then for any $u \in K$

$$\eta \sum_{t=1}^{T} [l_t(x_t) - l_t(u)] = D_{\Phi_0}(u, x_1) - D_{\Phi_T}(u, x_{T+1}) + \sum_{t=1}^{T} D_{\Phi_t}(x_t, x_{t+1}) \tag{5}$$

Aside: $\sum_{t=1}^{T} l_t(x_t) \leq \inf_{u \in K}[\sum l_t(u) + \eta^{-1} D_R(u, x_1)] + \eta \sum_{t=1}^{T} D_{\Phi_t}(x_t, x_{t+1})$

Proof. x_{t+1} minimizes Φ_t
$\nabla\Phi_t(x_{t+1}) = 0 \Rightarrow D_{\Phi_t}(u, x_{t+1}) = \Phi_t(u) - \Phi_t(x_{t+1})$
Moreover, $\Phi_t(u) = \Phi_{t-1}(u) + \eta l_t(u)$
Conditioning:

$$\begin{array}{rcl} (-) \quad \eta l_t(u) & = & D_{\Phi_t}(u, x_{t+1}) + \Phi_t(x_{t+1}) - \Phi_{t-1}(u) \\ (+) \quad \eta l_t(x_t) & = & D_{\Phi_t}(x_t, x_{t+1}) + \Phi_t(x_{t+1}) - \Phi_{t-1}(x_t) \\ ------------ & -- & ------------------------ \\ \eta[l_t(x_t) - l_t(u)] & = & D_{\Phi_t}(x_t, x_{t+1}) + D_{\Phi_{t-1}}(u, x_t) - D_{\Phi_t}(u, x_{t+1}) \end{array}$$

Sum over $t = 1 \cdots\cdots T$, we get the statement of the Lemma. □

3

Suppose $\nabla R(x_1) = 0, \quad \mathbb{R}^n = K$

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} [\eta l_t(x) + D_{\Phi_{t-1}}(x, x_t)] \tag{6}$$

Statement: two definitions eq.4 and eq.6 are equivalent.

$$\begin{array}{rcl} \eta l_t(x) & = & \Phi_t(x) - \Phi_{t-1}(x) \\ \eta l_t(x) + D_{\Phi_{t-1}}(x, x_t) & = & \Phi_t(x) - \Phi_{t-1}(x) + D_{\Phi_{t-1}}(x, x_t) \end{array}$$

Suppose that definitions are equivalent for $\tau \leq t$, x minimizes Φ_{t-1}.

$$\nabla_x D_{\Phi_{t-1}}(x, x_t) = \nabla_x \Phi_{t-1}(x) - \nabla_x \Phi_{t-1}(x_t)$$
$$\nabla\Phi_t(x_{t+1}) = \nabla\Phi_{t-1}(x_t) = \cdots\cdots = \nabla\mathbb{R}(x_1) = 0$$

thus $x_{t+1} = \arg\min_{x \in K} \Phi_t(x)$.

Suppose l_t's are linear functions abusing notation "$l_t \cdot x$".

Corollary 3.1. (1) $\eta(\sum l_t x_t - \sum l_t \cdot u) = D_R(u, x_1) - D_R(u, x_{t+1}) + \sum D_R(x_t, x_t - 1)$ for any $u \in \mathbb{R}^n$.
(2) $x_{t+1} = \nabla R^*(\nabla R(x_t) - \eta l_t)$

Proof. (1)$D_{\Phi_t} = D_R$ because $\Phi_t = R + \sum_{s=1}^{t} l_s$.
(2)

$$\begin{array}{rcl} x_t & \text{satisfies} & \eta \sum_{s=1}^{t-1} l_s + \nabla R(x_t) = 0 \\ x_{t+1} & \text{satisfies} & \eta \sum_{s=1}^{t} l_s + \nabla R(x_{t+1}) = 0 \\ \hline & & \eta l_t + \nabla R(x_{t+1}) - \nabla R(x_t) = 0 \\ & & x_{t+1} = \nabla R^*(\nabla R(x_t) - \eta l_t) \end{array}$$

□

Recall online gradient descent $x_{t+1} = x_t - \eta l_t$.

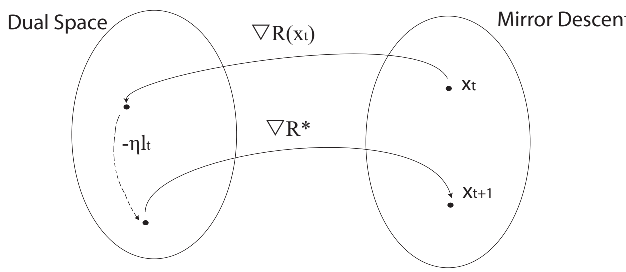

Figure 1: default

If $R = \frac{1}{2}||\cdot||^{\frac{1}{2}}$, $\quad \nabla R(x) = x, \quad \nabla R^*(x) = x$.
If $l_t(\cdot)$ are convex (but not necessary linear).

Lemma 3.2. If we choose $x_{t+1} = \arg\min_{x\in\mathbb{R}^n}[\eta\nabla l_t(x_t)^T x + D_R(x, x_t)]$ (or equivalently $x_{t+1} = \arg\min \eta\sum[\nabla l_t(x_t)^T x + R(x)]$. Then $\sum_{t=1}^{T}(l_t(x_t) - l_t(u)) \leq \eta^{-1}D_R(u, x_1) + \sum_{t=1}^{T} D_R(x_t, x_{t+1})$

Proof.

$$\sum_{t=1}^{T}[l_t(x_t) - l_t(u)] \leq \sum(\tilde{l}_t x_t - \tilde{l}_t u) \leq \cdots$$

□

4 Time-varying learning rate η_t

$$x_{t+1} = \arg\min \sum \eta_s l_s(x) + R(x)$$

Lemma 4.1. $K = \mathbb{R}^n$, Then for any $u \leq \mathbb{R}^n$,

$$\sum_{t=1} T[l_t(x_t) - l_t(u)] \leq \sum_{t=1} T\eta_t^{-1}[D_{\Phi_t}(x_t, x_{t+1}) + D_{\Phi_{t+1}}(u, x_t) - D_{\Phi_t}(u, x_{t+1})$$

Definition A function g is σ-strong convex with respect to R if all $x, y \in \mathbb{R}^n$, $g(x) \geq g(y) + \nabla g(y)^T(x-y) + \sigma/2 D_R(x,y)$

$$l_t(x_t) - l_t(u) \leq \tilde{l}_t(x_t) - \tilde{l}_t(u) - \frac{\sigma_t}{2} D_R(u, x_t)$$

Final result:

$$\begin{aligned} \sum [l_t(x_t) - l_t(u)] &\leq \sum [\tilde{l}_t(x_t) - \tilde{l}_t(u) - \frac{\sigma_t}{2} D_R(u, x_t)] \\ &\leq \sum_1^T \eta_t^{-1} D_R(x_t, x_{t+1}) + \sum_1^T (\eta_t^{-1} - \frac{\sigma_t}{2} \eta_{t-1}^{-1}) D_R(u, x_t) + (\eta_1^{-1} - \frac{\sigma_1}{2}) D_R(u, x_1) \end{aligned}$$

Sketch of the proof: If we take $\eta_t = (\frac{1}{2} \sum_{s=1} t\sigma_s)^{-1}$, we obtain $\sum [l_t(x_t) - l_t(u)] \leq \sum \eta_t^{-1} D_R(x_t, x_{t+1})$. If $R = \frac{1}{2}|| \cdot ||^2$, regret$\leq log(T)$.