## 5.7 Online mirror descent

Quadratic regularization is imposing a certain prior knowledge where there is a good $w_t$ in a small $L_2$ ball.

Now we develop a general way of obtaining regret bounds for general regularizers.

Goal: analyse FTRL for $\forall$ convex loss and regularizers.

[Algorithm 5] Online mirror descent (OMD)

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be the regularizer

Let $f_1, \cdots, f_T$ : sequence of losses.

On each iteration $t = 1, \cdots, T$, the learner chooses $w_t$ s.t.

$$w_t \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \{\psi(w) - w \cdot \theta_t\}. \tag{522}$$

where $z_t \in \partial f_t(w_t)$, $\theta_t = -\sum_{i=1}^{t-1} z_i$ (negative sum)

Technical Note: $S = \mathbb{R}^d$. And we can always fold constraints by setting $\psi(w) = \infty$ if $w$ violates the constraints.

[Examples of regularizers]

- Qudratic : $\psi(w) = \frac{1}{2\eta} \|w\|_2^2$
- Non-spherical quadratic : $\psi(w) = \frac{1}{2\eta} w^T A w$
- Entropic : $\psi(w) = \frac{1}{\eta} \sum_{i=1}^{d} w_j \log w_j$ if $w \in \Delta_d$

Difference: the slope when $w \to$ boundary.

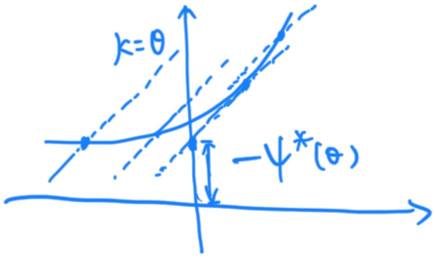
[Definition 27] Fenchel conjugate

The Fenchel conjugate of a function $\psi : \mathbb{R}^d \to \mathbb{R}$ is

$$\psi^*(\theta) := \sup_{w \in \mathbb{R}^d} \{w \cdot \theta - \psi(w)\}$$

Intuition: For scalars $w, \theta \in \mathbb{R}$, fixed $\theta$ (as slope), $-\psi^*(\theta)$

is the position where supporting hyperplane of $\psi$ with slope $\theta$ hit the vertical axis.




Useful facts:

① $\psi^*$ is always convex, since it's a supremum over a collection of linear function $\theta \mapsto w\cdot\theta - f(w)$

② $\psi^*(\theta) \geq w\cdot\theta - \psi(w)$ for all $w \in \mathbb{R}^d$ (French-Young inequality)

③ If $r(w) = a\psi(w)$ with $a > 0$, then $r^*(\theta) = a\psi^*(\theta/a)$

④ $\psi^{**} = \psi$ iff $\psi$ is convex (and low semi-continuous)

⑤ If $\psi$ is differentiable, then

$$\nabla\psi^*(\theta) = \arg\max_{w\in\mathbb{R}^d} \{w\cdot\theta - \psi(w)\}$$ {w is the gradient}

Mirroring

(i) OMD update (522): $w_t \in \arg\min_{w\in S} \{\psi(w) - w\cdot\theta_t\}$

we have $w_t = \nabla\psi^*(\theta)$ and $-\psi^*(\theta)$ is the corresponding value of regularized loss

(ii) Since $w_t$ attains $\sup\{\psi(w) - w\cdot\theta_t\}$, differentiate w.r.t. $w$ and we have $\theta_t = \nabla\psi(w_t)$

(iii) One-to-one mapping:

$$w_t = \nabla\psi^*(\theta_t), \quad \theta_t = \nabla\psi(w_t)$$

(iv) OMD updates: $\theta_{t+1} = \theta_t - z_t$, $\theta_1, \cdots, \theta_T$

by mirrored: $w_t = \nabla\psi^*(\theta_t)$, $w_1, \cdots, w_T$.

[Example 32] Quadratic

Let $\psi(w) = \frac{1}{2\eta}\|w\|_2^2$

Then $\psi^*(\theta) = \sup_{w \in S}\{w \cdot \theta - \frac{1}{2\eta}\|w\|_2^2\} = \frac{\eta}{2}\|\theta\|_2^2$

[Example 33] Entropic

Let $\psi(w) = \frac{1}{\eta}\sum_{j=1}^{d} w_j \log w_j$ for $w \in \Delta_d$

$\max\{w \cdot \theta - \psi(w)\}$ s.t. $w \in \Delta_d$.

Let $\theta_i - \frac{1}{\eta}(\log w_i + 1) = c$ , $w_i = e^{\eta\theta_i - \eta c - 1}$

$\sum w_i = 1$ ,

$\Rightarrow w_i = \frac{e^{\eta\theta_i}}{\sum e^{\eta\theta_j}}$

$\Rightarrow \psi^*(w) = \sum_i \frac{\theta_i e^{\eta\theta_i}}{\sum_j e^{\eta\theta_j}} - \frac{1}{\eta}\sum_i \frac{e^{\eta\theta_i}}{\sum e^{\eta\theta_j}}[\eta\theta_i - \log \sum e^{\eta\theta_j}]$

$\Rightarrow \psi^*(w) = \frac{1}{\eta}\log(\sum_{j=1}^{d} e^{\eta\theta_j})$