Online Learning and Online Convex Optimization

Chapter 2 in Shai Shalev Shwartz / Online Learning and Online convex Optimization

Outline

Online Convex Optimization (OCO)

Algorithm

Input: A convex set S

For $t = 1, 2, \ldots$

- Predict a vector $w_t \in S$
- Receive a convex loss function $f_t : S \to \mathbb{R}$
- Suffer loss $f_t(w_t)$

Regret Definition

Regret of the Algorithm:

$$\text{Regret}_T(u) = \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u). \tag{1}$$

Regret relative to a set of vectors U:

$$\text{Regret}_T(U) = \max_{u \in U} \text{Regret}_T(u). \tag{2}$$

Follow-the-Leader Algorithm

FTL Strategy

At round t, select:

$$w_t = \text{argmin}_{w \in S} \sum_{i=1}^{t-1} f_i(w)$$

- Natural approach: Choose best performer on past data
- Simple but can be unstable
- Requires solving optimization problem each round

FTL Regret Analysis

Theorem (Lemma 2.1)

For any $u \in S$:

$$Regret_T(\mathrm{u}) = \sum_{t=1}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{u})) \leq \sum_{t=1}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1}))\,.$$

proof

Step 1: Equivalent to

$$\sum_{t=1}^{T} f_t(\mathrm{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathrm{u})$$

Step 2: By induction on T:

- Base case: $T = 1$ trivial as $f_1(w_1) - f_1(u) \leq 0$
- Inductive step: Assume holds for $T - 1$, then

$$\begin{aligned}\sum_{t=1}^{T}[f_t(w_t) - f_t(u)] \\ &= \underbrace{\sum_{t=1}^{T-1}[f_t(w_t) - f_t(u)]}_{\leq \sum_{t=1}^{T-1}[f_t(w_t) - f_t(w_{t+1})]} + [f_T(w_T) - f_T(u)] \\ &\leq \sum_{t=1}^{T}[f_t(w_t) - f_t(w_{t+1})]\end{aligned}$$

using $w_{T+1} = \text{argmin}_w \sum_{t=1}^{T} f_t(w)$

FTL for Quadratic Optimization

For $f_t(w) = \frac{1}{2}\|w - z_t\|_2^2$:

- FTL update: $w_t = \frac{1}{t-1}\sum_{i=1}^{t-1} z_i$
- Regret bound: $O(\log T)$

Regret Calculation for quadratic optimization.

$$\begin{aligned}
\text{Regret}_T(u) &\leq \sum_{t=1}^{T} \frac{1}{t}\|w_t - z_t\|^2 \\
&\leq \sum_{t=1}^{T} \frac{(2L)^2}{t} = 4L^2(\log T + 1)
\end{aligned}$$

where $L = \max_t \|z_t\|$ $\square$

Failure of follow the leader

$f_t(w) = w \cdot z$:

- $$z_t = \begin{cases} -0.5 & \text{if } t = 1 \\ 1 & \text{if } t \text{ is even} \\ -1 & \text{if } t > 1 \text{ and } t \text{ is odd} \end{cases}$$
- $w_t = -1, 1, -1, 1, \ldots$
- Cumulative loss is T.
- Cumulative loss of 0 is 0
- Regret is T.
- **Reason:** prediction is unstable
- We need to regularize.
- $R(W)$ penalizes vectors which are large.

Follow-the-Regularized-Leader (FTRL)

$$\forall t, \quad \mathrm{w}_t = \arg\min_{\mathrm{w} \in S} \sum_{i=1}^{t-1} f_i(\mathrm{w}) + R(\mathrm{w})$$

- For bad case above: $w_t = 0, 0, 0, 0, \ldots$
- Each step requires solving a minimization problem.

Lemma 2.3: Follow-the-Regularized-Leader

Lemma 2.3. Let $\mathrm{w}_1, \mathrm{w}_2, \ldots$ be the sequence of vectors produced by FoReL. Then, for all $\mathrm{u} \in S$ we have:

$$\sum_{t=1}^{T}(f_t(\mathrm{w}_t) - f_t(\mathrm{u})) \leq R(\mathrm{u}) - R(\mathrm{w}_1) + \sum_{t=1}^{T}(f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1})).$$

Proof of Lemma 2.3

Proof. Observe that running FoReL on $f_1, \ldots, f_T$ is equivalent to running FTL on $f_0, f_1, \ldots, f_T$ where $f_0 = R$. Using Lemma 2.1, we obtain:

$$\sum_{t=0}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{u})) \leq \sum_{t=0}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1})).$$

Rearranging the above and using $f_0 = R$, we conclude our proof.
$\square$

FTRL Regret Bound for linear functions

For linear $f_t(w) = \langle w, z_t \rangle$ and $R(w) = \frac{1}{2\eta}\|w\|_2^2$
Update rule $w_{t+1} = w_t - \eta z_t$ Then, for all u we have

$$\mathrm{Regret}_T(\mathrm{u}) \leq \frac{1}{2\eta}\|\mathrm{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathrm{z}_t\|_2^2.$$

Choice of η and Final Bound for linear functions

Tunings:

- Define the set $U = \{\mathbf{u} : \|\mathbf{u}\| \leq B\}$.
- Assume that
$$\frac{1}{T}\sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2 \leq L^2.$$
- Set $\eta = \frac{B}{L\sqrt{2T}}$.

Conclusion:

$$\text{Regret}_T(U) \leq BL\sqrt{2T}.$$

From linear functions to Online Gradient Descent

Example (OGD from FTRL)

Consider the OCO setup where the functions $f_1, f_2, \ldots$ are differentiable.
Let η be the learning rate.

$$w_{t+1} = w_t - \eta z_t, \;\; z_t = \nabla f_t(w_t)$$

Identical to FTRL with regularization: $R(w) = \frac{1}{2\eta}\|w\|_2^2$

Regret bound on OGD: From FTRL theorem:

$$\begin{aligned} \text{Regret} &\leq \frac{\|u\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|z_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \eta T L^2 \end{aligned}$$

Regret Bound for OGD

If we further assume that each f_t is L_t-Lipschitz with respect to $\|\cdot\|_2$, and let L be such that

$$\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2.$$

Then, for all u, the regret of OGD satisfies

$$\mathrm{Regret}_T(\mathrm{u}) \leq \frac{1}{2\eta}\|\mathrm{u}\|_2^2 + \eta T L^2.$$

Bounding the norm of $\mathbf{u}$

In particular, if $U = \{\mathbf{u} : \|\mathbf{u}\|_2 \leq B\}$ and $\eta = \frac{B}{L\sqrt{2T}}$ then

$$\text{Regret}_T(U) \leq BL\sqrt{2T}.$$

Practical Considerations

Doubling Trick

- Removes need to know time horizon T
- Divide time into epochs $2^m, 2^{m+1} - 1$
- Regret increases by constant factor:

$$\sum_{m=0}^{\log T} \sqrt{2^m} = O(\sqrt{T})$$

Example (Optimal η)

Setting $\eta = \frac{B}{L}\sqrt{\frac{2}{T}}$ gives:

$$BL\sqrt{2T}$$

Definition 2.4: Strong Convexity

Strong Convexity

A function $f : S \to \mathbb{R}$ is σ-strongly convex over S with respect to a norm $\|\cdot\|$ if for any $\mathrm{w} \in S$ we have:

$$\forall \mathrm{z} \in \partial f(\mathrm{w}), \quad \forall \mathrm{u} \in S, \quad f(\mathrm{u}) \geq f(\mathrm{w}) + \langle \mathrm{z}, \mathrm{u} - \mathrm{w} \rangle + \frac{\sigma}{2}\|\mathrm{u} - \mathrm{w}\|^2.$$

Lemma 2.8: Strong Convexity implication

Lemma 2.8

Let S be a nonempty convex set. Let $f : S \to \mathbb{R}$ be a σ-strongly convex function over S with respect to a norm $\|\cdot\|$. Let:

$$\mathrm{w} = \arg\min_{\mathrm{v} \in S} f(\mathrm{v}).$$

Then, for all $\mathrm{u} \in S$, we have:

$$f(\mathrm{u}) - f(\mathrm{w}) \geq \frac{\sigma}{2}\|\mathrm{u} - \mathrm{w}\|^2.$$

Strong Convexity Condition

If R is twice differentiable, then it is easy to verify that a sufficient condition for strong convexity of R is that for all w, x,

$$\langle \nabla^2 R(\mathrm{w})\mathrm{x}, \mathrm{x} \rangle \geq \sigma \|\mathrm{x}\|^2,$$

where $\nabla^2 R(\mathrm{w})$ is the Hessian matrix of R at w, namely, the matrix of second-order partial derivatives of R at w [39, Lemma 14].

Example 2.4: Euclidean Regularization

The function

$$R(w) = \frac{1}{2}\|w\|_2^2$$

is 1-strongly-convex with respect to the ℓ_2 norm over $\mathbb{R}^d$. To see this, simply note that the Hessian of R at any w is the identity matrix.

Example 2.5: Entropic Regularization

The function

$$R(\mathrm{w}) = \sum_{i=1}^{d} w[i] \log(w[i])$$

is $\frac{1}{B}$-strongly-convex with respect to the ℓ_1 norm over the set

$$S = \{\mathrm{w} \in \mathbb{R}^d : \mathrm{w} > 0 \wedge \|\mathrm{w}\|_1 \leq B\}.$$

In particular, R is 1-strongly-convex over the probability simplex, which is the set of positive vectors whose elements sum to 1.

Proof of strong convexity for Entropic Regularization

$\frac{\partial^2}{\partial w[i]^2} w[i] \log w[i] = \frac{1}{w[i]}$

$$\langle \nabla^2 R(\mathrm{w})\mathrm{x}, \mathrm{x} \rangle = \sum_i \frac{x[i]^2}{w[i]}$$

$$= \frac{1}{\|\mathrm{w}\|_1} \left(\sum_i w[i] \right) \left(\sum_i \frac{x[i]^2}{w[i]} \right)$$

$$\geq \frac{1}{\|\mathrm{w}\|_1} \left(\sum_i \sqrt{w[i]} \frac{x[i]}{\sqrt{w[i]}} \right)^2 = \frac{\|\mathrm{x}\|_1^2}{\|\mathrm{w}\|_1},$$

where the inequality follows from Cauchy–Schwarz inequality.

Single Step of FTRL with Strong Convexity

Let

$$R : S \to \mathbb{R}$$

be a σ-strongly-convex function over S with respect to a norm $\|\cdot\|$. Let $\mathrm{w}_1, \mathrm{w}_2, \ldots$ be the predictions of the FoReL algorithm. Then, for all t, if f_t is L_t-Lipschitz with respect to $\|\cdot\|$, then:

$$f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1}) \leq L_t \|\mathrm{w}_t - \mathrm{w}_{t+1}\| \leq \frac{L_t^2}{\sigma}.$$

Proof (Single Step of FTRL with Strong Convexity)

For all t let

$$F_t(\mathrm{w}) = \sum_{i=1}^{t-1} f_i(\mathrm{w}) + R(\mathrm{w})$$

and note that the FoReL rule is

$$\mathrm{w}_t = \arg\min_{\mathrm{w}\in S} F_t(\mathrm{w}).$$

Note also that F_t is σ-strongly-convex since the addition of a convex function to a strongly convex function keeps the strong convexity property. Therefore, Lemma 2.8 implies that:

$$F_t(\mathrm{w}_{t+1}) \geq F_t(\mathrm{w}_t) + \frac{\sigma}{2}\|\mathrm{w}_t - \mathrm{w}_{t+1}\|^2.$$

Continuing the Proof (Single Step of FTRL with Strong Convexity)

Repeating the same argument for F_{t+1} and its minimizer w_{t+1}, we get:

$$F_{t+1}(\mathrm{w}_t) \geq F_{t+1}(\mathrm{w}_{t+1}) + \frac{\sigma}{2}\|\mathrm{w}_t - \mathrm{w}_{t+1}\|^2.$$

Taking the difference between the last two inequalities and rearranging, we obtain:

$$\sigma\|\mathrm{w}_t - \mathrm{w}_{t+1}\|^2 \leq f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1}). \quad (2.7)$$

Final Steps (Single Step of FTRL with Strong Convexity)

Next, using the Lipschitzness of f_t, we get that:

$$f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1}) \leq L_t \|\mathrm{w}_t - \mathrm{w}_{t+1}\|.$$

Combining with Equation (2.7) and rearranging, we get:

$$\|\mathrm{w}_t - \mathrm{w}_{t+1}\| \leq L/\sigma.$$

Together with the above, we conclude our proof. ■

Main theorem regarding σ-strongly convex regularization functions

Let $f_1, \ldots, f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to some norm $\|\cdot\|$. Let L be such that

$$\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2.$$

Assume that FoReL is run on the sequence with a regularization function which is σ-strongly-convex with respect to the same norm. Then, for all $\mathrm{u} \in S$,

$$\mathrm{Regret}_T(\mathrm{u}) \leq R(\mathrm{u}) - \min_{\mathrm{v}\in S} R(\mathrm{v}) + \frac{TL^2}{\sigma}.$$

Corollary for l_2 regularization

Let $f_1, \ldots, f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to $\|\cdot\|_2$. Let L be such that

$$\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2.$$

Assume that FoReL is run on the sequence with the regularization function

$$R(\mathrm{w}) = \frac{1}{2\eta}\|\mathrm{w}\|_2^2.$$

Then, for all u,

$$\mathrm{Regret}_T(\mathrm{u}) \leq \frac{1}{2\eta}\|\mathrm{u}\|_2^2 + \eta T L^2.$$

Applications to expert advice

- Distribution w_t
- Action Losses: $x_t \in [0,1]^d$
- Algorithm Loss: $\langle x_t, w_t \rangle$
- We want to bound regret.
- we will compare l_2 regularization with Entropic Regularization.

Experts using l_2 regularization (1)

S be a convex set and consider running FoReL with the regularization function:

$$R(\mathrm{w}) = \begin{cases} \frac{1}{2\eta}\|\mathrm{w}\|_2^2 & \text{if } \mathrm{w} \in S \\ \infty & \text{if } \mathrm{w} \notin S \end{cases}$$

Where S us the d dimensional simplex.
Then, for all $\mathrm{u} \in S$,

$$\mathrm{Regret}_T(\mathrm{u}) \leq \frac{1}{2\eta}\|\mathrm{u}\|_2^2 + \eta T L^2.$$

Experts using l_2 regularization (2)

If

$$B \geq \max_{\mathbf{u} \in S} \|\mathbf{u}\|_2$$

Setting

$$B = 1;\ L = \sqrt{d};\ \eta = \frac{B}{L\sqrt{2T}} = \frac{1}{\sqrt{2dT}}$$

then,

$$\mathrm{Regret}_T(S) \leq \sqrt{2dT}.$$

Entropic Regularization

Let $f_1, \ldots, f_T$ be a sequence of convex functions such that f_t is L_t-Lipschitz with respect to $\|\cdot\|_1$. Let L be such that $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$. Assume that FoReL is run on the sequence with the regularization function

$$R(\mathrm{w}) = \frac{1}{\eta}\sum_i w[i]\log(w[i])$$

and with the set

$$S = \{\mathrm{w} : \|\mathrm{w}\|_1 = B \wedge \mathrm{w} > 0\} \subset \mathbb{R}^d.$$

Then,

$$\mathrm{Regret}_T(S) \leq \frac{B\log(d)}{\eta} + \eta BTL^2.$$

In particular, setting $\eta = \frac{\sqrt{\log d}}{L\sqrt{2T}}$ yields

$$\mathrm{Regret}_T(S) \leq BL\sqrt{2\log(d)T}.$$

Entropic regularization for Experts

The Entropic regularization is strongly convex with respect to the ℓ_1 norm, and therefore the Lipschitzness requirement of the loss functions is also with respect to the ℓ_1-norm.
For linear functions,

$$f_t(\mathrm{w}) = \langle \mathrm{w}, \mathrm{x_t} \rangle,$$

we have by Hölder's inequality that,

$$|f_t(\mathrm{w}) - f_t(\mathrm{u})| = |\langle \mathrm{w} - \mathrm{u}, \mathrm{x_t} \rangle| \leq \|\mathrm{w} - \mathrm{u}\|_1 \|\mathrm{x_t}\|_\infty .$$

Therefore, the Lipschitz parameter grows with the ℓ_∞ norm of $\mathrm{x_t}$ rather than the ℓ_2 norm of $\mathrm{x_t}$.
expert advice: $B = 1$ and $L = 1$), we obtain the regret bound of

$$\sqrt{2 \log(d) T}.$$

Comparison between regularizations

- entropic regularization vs. ℓ_2 regularization.
- $\log d$ vs $\sqrt{d}$
- L: $\|x_t\|_\infty \geq \|x_t\|_2$ Liphsitz condition carries heavier penalty with entropic regularization.
- B: $\|u\|_1 \leq \|u\|_2$ Comparator length carries heavier penalty with l_2 norm.

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\text{w}) = L_{A,t} - L_t(\text{w})$
- Regret_t = State of prediction algorithm at time t
- Potential/Regularizer: $R(\text{Regret})$ Quantifies badness of the state.
- A state is bad if adversary can force high regret in the future.
- Choose prediction so that $R(\text{Regret}_{t+1}) - R(\text{Regret}_t) + \text{w}_t \cdot \ell_t$ is small for all possible ℓ_t
- $\text{w}_t = \nabla R(\text{Regret}_t)$ is a good choice.
- For finite number of experts, Regret_t is finite dimensional and we can compute w_t explicitly.
- Here, $\text{Regret} = \{R(\text{w})\}_{\text{w} \in \mathbb{R}^d}$ is uncountably infinite.
- If Experts correspond to exponential distributions and loss is log loss- we can use conjugate priors. (recall: biased coins).
- We need a new trick to compute $\text{w}_t = \nabla R(\text{Regret}_t)$ efficiently.

FoReL Update Rule for linear cost function

Define $z_{1:t} = \sum_{i=1}^{t} z_i$, the FoReL update rule can be written as

$$
\begin{aligned}
w_{t+1} &= \arg\min_{w} R(w) + \sum_{i=1}^{t} \langle w, z_i \rangle \\
&= \arg\min_{w} R(w) + \langle w, z_{1:t} \rangle \\
&= \arg\max_{w} \langle w, -z_{1:t} \rangle - R(w).
\end{aligned}
$$

Mirror Descent Update for linear functions

Update rule

$$w_{t+1} = \arg\max_{w} \langle w, -z_{1:t} \rangle - R(w).$$

Link Function:

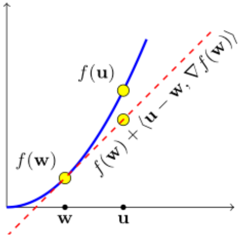
$$g(\theta) = \arg\max_{w} \langle w, \theta \rangle - R(w),$$

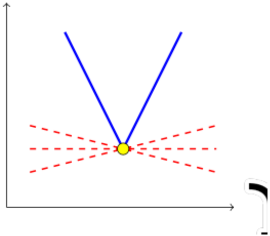
Update rule can be re-written as

1. $\theta_0 = 0$
2. $\theta_{t+1} = \theta_t - z_t$
3. $w_{t+1} = g(\theta_{t+1})$

Sub-Gradients

- we can reduce general convex to linear using the gradient.
- What can we do if $f(x)$ is convex but not differentiable at x?
- Use the sub-gradients at $x \doteq \partial f(x)$: the set of linear functions such that $l(x) = \langle w, x \rangle + o$ such that $\forall y, l(y) \leq f(x)$ and $l(x) = f(x)$
- if gradient $\nabla f(x)$ exists, then $\partial f(x) = \{\nabla f(x)\}$





Example Generalized Online Gradient Descent

Consider the ℓ_2 setup where the functions $f_1, f_2, \ldots$ are convex (but not necessarily differentiable). Let η be the learning rate.

$$w_{t+1} = w_t - \eta z_t, \;\; z_t \in \partial f_t(w_t)$$

Identical to FTRL with regularization: $R(w) = \frac{1}{2\eta}\|w\|_2^2$
Regret bound on OGD: From FTRL theorem:

$$\begin{aligned}\text{Regret} &\leq \frac{\|u\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|z_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \eta T L^2\end{aligned}$$

Online Mirror Descent (OMD)

parameter: a link function $g : \mathbb{R}^d \to S$
initialize: $\theta_1 = 0$
for $t = 1, 2, \ldots$

- **predict** $\mathrm{w}_t = g(\theta_t)$
- **update** $\theta_{t+1} = \theta_t - \mathrm{z}_t$ where $\mathrm{z}_t \in \partial f_t(\mathrm{w}_t)$

Duality

- OMD can be analyzed using elementary tools.
- Using Duality Gives better intuition, more general analysis, tighter bounds.

Dual Vector Spaces

- V is a vector space, with a norm $\|v\|$
- U is the set of all linear mappings from V to V
- The norm of $u \in U$ is defined as

$$\|u\|^* = \max_{v \in V} \frac{\|u(v)\|}{\|v\|}$$

- V is equivalent to the set of all linear mappings from U to U.
- U and V are dual vector spaces, with dual norms.

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- L_2 norm: $\sqrt{\sum_{i=1}^n x_i^2}$
- L_1 norm: $\sum_{i=1}^n |x_i|$
- L_∞ norm: $\max_i |x_i|$
- L_p norm: $\left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}$
- L_p, L_q are dual norms if $p, q \geq 1$, and $\frac{1}{p} + \frac{1}{q} = 1$
- L_1, L_∞ are dual.
- L_2 is self-dual.

Fenchel Duality

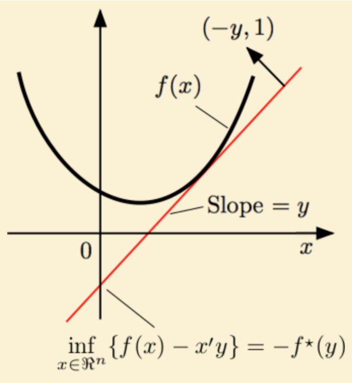
- Suppose $F : A \to \mathbb{R}$ is a convex function over a convex set $A \subseteq \mathbb{R}^n$.
- The dual function to F is

$$F^*(\mathrm{u}) = \sup_{\mathrm{v} \in A} \left(\mathrm{u} \cdot \mathrm{v} - F(\mathrm{v})\right)$$

- Fenchel duality Reduces to Legendre duality for differentiable functions

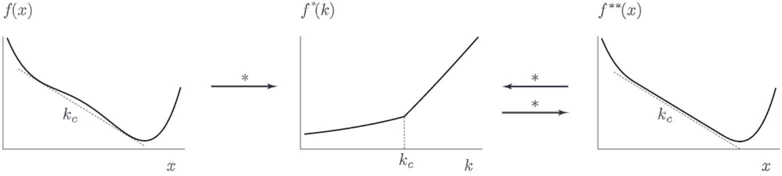
Visualization of the Febchel Dual

- $x, y\mathbb{R}$
- $f^*(y) = \sup_{x\in\mathbb{R}}(xy - f(x))$
- $-f^*(y) = \inf_{x\in\mathbb{R}}(f(x) - xy)$



Dual of Dual

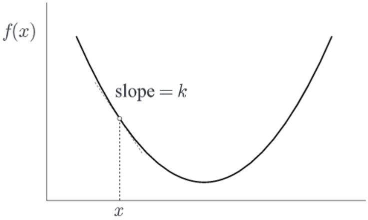
- The dual of any function is convex.
- if F is convex then $F^{**} = F$

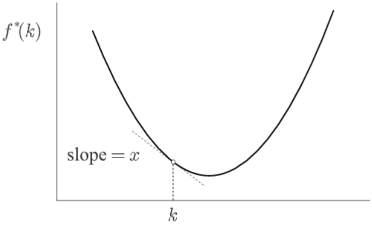


Gradient Duality

- If the gradient of f at x is k then the gradient of f^* at k is x
- In general:

$$\nabla F^* = (\nabla F)^{-1}$$



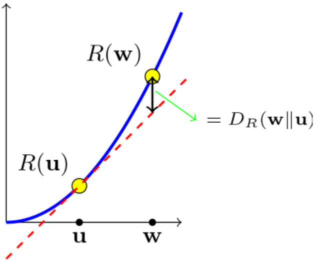


Example: Exponential Potential

- Potential: $F(\mathrm{u}) = \sum_{i=1}^{d} e^{u_i}$
- Gradient: $\nabla F(\mathrm{u})_i = e^{u_i}$ or $\nabla F(\mathrm{u}) = F(\mathrm{u})$.
- Dual: $F^*(\mathrm{v}) = \sum_{i=1}^{d} v_i(\ln v_i - 1)$
- Gradient of dual: $\nabla F^*(\mathrm{v})_i = \ln v_i$
- Note $(\nabla F)^{-1} = \nabla F^*$

Bregman Divergence

- $R(x)$ is convex and differentiable.
- $D_R(w||u) = R(w) - (R(u) + \langle \nabla R(u), (w - u) \rangle)$



Fenchel and Bregman

- F: strictly convex with continuous first derivative.
- F^* is the Fenchel Dual of F
- D_F, D_{F^*} Bregman divergences wrt F, F^*
- $\mathrm{u}' = \nabla F(\mathrm{u})$ and $\mathrm{v}' = \nabla F(\mathrm{v})$
- $D_F(\mathrm{u}, \mathrm{v}) = D_{F^*}(\mathrm{u}', \mathrm{v}')$

Mirror Descent

- Gradient descent in dual space $\theta_t = \theta_{t-1} - \lambda \nabla \ell_t(\theta_{t-1})$
- Using duality can be rewritten as

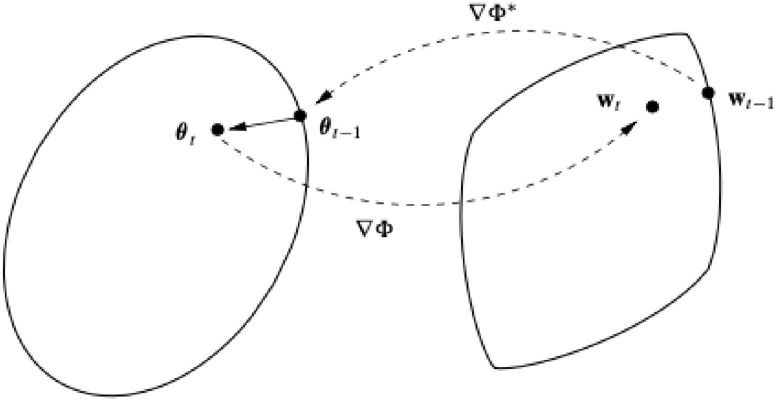
$$\nabla R^*(\mathrm{w}_t) = \nabla R^*(\mathrm{w}_{t-1}) - \lambda \nabla \ell_t(\mathrm{w}_{t-1})$$

- As ∇R is the inverse of ∇R^* we get

$$\mathrm{w}_t = \nabla R(\nabla R^*(\mathrm{w}_{t-1}) - \lambda \nabla \ell_t(\mathrm{w}_{t-1}))$$

A picture of mirror descent

$$w_t = \nabla R(\nabla R^*(w_{t-1}) - \lambda \nabla \ell_t(w_{t-1}))$$



Intuition

- u should balance minimizing the loss from observing same example again and divergence between u and w_{t-1}
- Exact Goal: $\min_{u \in \mathbb{R}^d} [D_{\phi^*}(u, w_{t-1}) - \lambda \nabla \ell_t(u)]$
- Taylor order one approximation: $\min_{u \in \mathbb{R}^d} [F(u)]$ where $F(u) = D_{\phi^*}(u, w_{t-1}) - \lambda[\ell_t(w_{t-1}) + (u - w_{t-1}) \nabla \ell_t(w_{t-1})]$
- Assuming everything is differrentiable and convex, $\nabla_u F[u] = 0$ yields: $\nabla R^*(w_t) = \nabla R^*(w_{t-1}) - \lambda \nabla \ell_t(w_{t-1})$
- Equivelently: $w_t = \nabla R(\nabla R^*(w_{t-1}) - \lambda \nabla \ell_t(w_{t-1}))$

Theorem

- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a regular loss function if it is convex and non-negative.
- Regret: $\text{Regret}_t(\mathrm{u}) = L_{A,t} - L_t(\mathrm{u})$
- Theorem: For all example sequences $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_T, y_T)$, any initial vector $\mathrm{w}_0 \in \mathbb{R}^d$. all learning rates $\lambda > 0$ and all $\mathrm{u} \in \mathbb{R}^d$:

$$\text{Regret}_T(\mathrm{u}) \leq \frac{1}{\lambda} D_{R^*}(\mathrm{u}, \mathrm{w}_0) + \frac{1}{\lambda} \sum_{t=1}^{T} D_{R^*}(\mathrm{w}_{t-1}, \mathrm{w}_t)$$

- $D_{R^*}(\mathrm{u}, \mathrm{w}_0)$ penalizes for the length of the comparator.
- $D_{R^*}(\mathrm{w}_{t-1}, \mathrm{w}_t)$ penalizes large changes in w_t.

Polynomial Potential

- Potential: $R_p(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_p^2 = \frac{1}{2}\left(\sum_{i=1}^{d} u_i^p\right)^{2/p}$
- Dual Potential $R_p^* = R_q$ Where $\frac{1}{p} + \frac{1}{q} = 1$
- Euclidean norm: $q = p = 2$
- Suppose the sequence of examples $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_T, y_T)$ satisfies $\|\mathrm{x}_t\|_p \leq X_p$ for all $1 \leq t \leq T$
- Suppose we use the dual descend algorithm for the potential function R_p and the learning rate $\lambda = \frac{2\epsilon}{(p-1)X_p^2}$ for some $0 < \epsilon < 1$
- Loss Bound:
  $L_{A,T} \leq \frac{L_T(\mathrm{u})}{1-\epsilon} + \frac{\|\mathrm{u}\|_q^2}{\epsilon(1-\epsilon)} \times \frac{(p-1)X_p^2}{4}$

Exponential Potential

- Potential: $R(\mathrm{u}) = \sum_{i=1}^{d} e^{u_i}$
- Dual Potential $R^*(\mathrm{u}) = \sum_{i=1}^{d} u_i(\ln u_i - 1)$
- Euclidean norm: $q = p = 2$
- Suppose the sequence of examples $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_T, y_T)$ satisfies $\|\mathrm{x}_t\|_\infty \leq X_p$ for all $1 \leq t \leq T$
- Suppose we use the dual descend algorithm for the exponential potential function R and the learning rate $\lambda = \frac{2\epsilon}{X_\infty^2}$ for some $0 < \epsilon < 1$
- Loss Bound:
  $L_{A,T} \leq \frac{L_T(\mathrm{u})}{1-\epsilon} + \frac{X_\infty^2 \ln d}{2\epsilon(1-\epsilon)}$

Lemma 2.20: Regret Bound for OMD

Lemma 2.20. Suppose that OMD is run with a link function $g = \nabla R^*$. Then, its regret is upper bounded by:

$$\sum_{t=1}^{T} \langle \mathrm{w}_t - \mathrm{u}, \mathrm{z}_t \rangle \leq R(\mathrm{u}) - R(\mathrm{w}_1) + \sum_{t=1}^{T} D_{R^*}(-\mathrm{z}_{1:t} \| - \mathrm{z}_{1:t-1}).$$

Furthermore, equality holds for the vector u that minimizes $R(\mathrm{u}) + \sum_t \langle \mathrm{u}, \mathrm{z}_t \rangle$.

Proof: Step 1 - Fenchel–Young Inequality

Using the **Fenchel–Young inequality**, we have:

$$R(\mathrm{u}) + \sum_{t=1}^{T} \langle \mathrm{u}, \mathrm{z}_t \rangle = R(\mathrm{u}) - \langle \mathrm{u}, -\mathrm{z}_{1:T} \rangle \geq -R^*(-\mathrm{z}_{1:T}).$$

Equality holds for u that maximizes $\langle \mathrm{u}, -\mathrm{z}_{1:T} \rangle - R(\mathrm{u})$, hence minimizing $R(\mathrm{u}) + \langle \mathrm{u}, \mathrm{z}_{1:T} \rangle$.

Proof: Step 2 - Bregman Divergence

Since $\mathrm{w}_t = \nabla R^*(-\mathrm{z}_{1:t-1})$ and using the definition of the Bregman divergence, we rewrite:

$$-R^*(-\mathrm{z}_{1:T}) = -R^*(0) - \sum_{t=1}^{T} \left(R^*(-\mathrm{z}_{1:t}) - R^*(-\mathrm{z}_{1:t-1}) \right).$$

Rearranging, we get:

$$= -R^*(0) + \sum_{t=1}^{T} \left(\langle \mathrm{w}_t, \mathrm{z}_t \rangle - D_{R^*}(-\mathrm{z}_{1:t} \| - \mathrm{z}_{1:t-1}) \right).$$

Conclusion

Note: Since

$$R^*(0) = \max_{w} \langle 0, w \rangle - R(w) = -\min_{w} R(w) = -R(w_1),$$

combining all the above, we conclude the proof. □