Dual Descent

February 19, 2025

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

Review: Property of FoRel Algorithm

Lemma 2.3:
Let $w_1, w_2, \ldots$ be the sequence of vectors produced by the FoReL algorithm. Then, for all $u \in S$, we have:

$$\sum_{t=1}^{T}(f_t(w_t) - f_t(u)) \leq R(u) - R(w_1) + \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1}))$$

Review: One step of Gradient Descent using strongly convex regularizer

Lemma 2.10:
Let $R : S \to \mathbb{R}$ be a σ-strongly-convex function over S with respect to a norm $\|\cdot\|$. Let $w_1, w_2, \ldots$ be the predictions of the FoReL algorithm. Then, for all t, if f_t is L_t-Lipschitz with respect to $\|\cdot\|$, we have:

$$f_t(w_t) - f_t(w_{t+1}) \leq L_t\|w_t - w_{t+1}\| \leq \frac{L_t^2}{\sigma}$$

Main Theorem regarding FoReL using stongly convex regularizer

Let $f_1, \dots, f_T$ be a sequence of convex functions with the following conditions:

- f_t is L_t-Lipschitz with respect to some norm $\| \cdot \|$.

Then, for all $u \in S$,

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

Main Theorem regarding FoReL using stongly convex regularizer

Let $f_1, \ldots, f_T$ be a sequence of convex functions with the following conditions:

- f_t is L_t-Lipschitz with respect to some norm $\|\cdot\|$.
- L satisfies $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$.

Then, for all $u \in S$,

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

Main Theorem regarding FoReL using stongly convex regularizer

Let $f_1, \ldots, f_T$ be a sequence of convex functions with the following conditions:

- f_t is L_t-Lipschitz with respect to some norm $\|\cdot\|$.
- L satisfies $\frac{1}{T}\sum_{t=1}^{T} L_t^2 \leq L^2$.
- FoReL is run on the sequence with a regularization function that is σ-strongly-convex with respect to the same norm.

Then, for all $u \in S$,

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

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\text{w}) = L_{A,t} - L_t(\text{w})$

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 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.

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\mathrm{w}) = L_{A,t} - L_t(\mathrm{w})$
- 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.

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

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\mathrm{w}) = L_{A,t} - L_t(\mathrm{w})$
- 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 $R(\text{Regret}_{t+1}) - R(\text{Regret}_t) + \mathrm{w}_t \cdot \ell_t$ is small for all possible ℓ_t
- $\mathrm{w}_t = \nabla R(\text{Regret}_t)$ is a good choice.

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\mathrm{w}) = L_{A,t} - L_t(\mathrm{w})$
- 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 $R(\text{Regret}_{t+1}) - R(\text{Regret}_t) + \mathrm{w}_t \cdot \ell_t$ is small for all possible ℓ_t
- $\mathrm{w}_t = \nabla R(\text{Regret}_t)$ is a good choice.
- For finite number of experts, Regret_t is finite dimensional and we can compute w_t explicitly.

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.

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\mathrm{w}) = L_{A,t} - L_t(\mathrm{w})$
- 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 $R(\text{Regret}_{t+1}) - R(\text{Regret}_t) + \mathrm{w}_t \cdot \ell_t$ is small for all possible ℓ_t
- $\mathrm{w}_t = \nabla R(\text{Regret}_t)$ is a good choice.
- For finite number of experts, Regret_t is finite dimensional and we can compute w_t explicitly.
- Here, $\text{Regret} = \{R(\mathrm{w})\}_{\mathrm{w} \in \mathbb{R}^d}$ is uncountably infinite.
- If Experts correspond to exponential distributions and loss is log loss- we can use conjugate priors. (recall: biased coins).

Potential based gradient Descent

- Regret_t = Regret vector $\text{Regret}_t(\mathrm{w}) = L_{A,t} - L_t(\mathrm{w})$
- 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 $R(\text{Regret}_{t+1}) - R(\text{Regret}_t) + \mathrm{w}_t \cdot \ell_t$ is small for all possible ℓ_t
- $\mathrm{w}_t = \nabla R(\text{Regret}_t)$ is a good choice.
- For finite number of experts, Regret_t is finite dimensional and we can compute w_t explicitly.
- Here, $\text{Regret} = \{R(\mathrm{w})\}_{\mathrm{w} \in \mathbb{R}^d}$ 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 $\mathrm{w}_t = \nabla R(\text{Regret}_t)$ 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. $\theta_0 = 0$

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. $\theta_{t+1} = \theta_t - z_t$

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. $w_{t+1} = g(\theta_{t+1})$

Identical update to FTRL for linear loss functions.
What about general convex loss functions?

Sub-Gradients

- we can reduce general convex to linear using the gradient.

Sub-Gradients

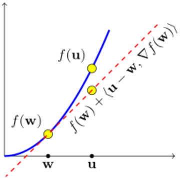
- we can reduce general convex to linear using the gradient.
- What can we do if $f(x)$ is convex but not differentiable at x?

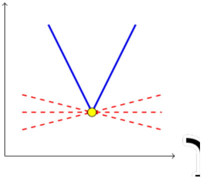
Sub-Gradients

- 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 $l(x) = f(x)$

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.

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

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

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

Gradient based Online Mirror Descent (OMD)

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

- **predict** $w_t = g(\theta_t)$

Dual Decent: Instead of minimizing f, minimize ∇f.
Convexity implies equivalence of goals.

Gradient based 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 $\mathrm{z}_t \in \partial f_t(\mathrm{w}_t)$

Dual Decent: Instead of minimizing f, minimize ∇f.
Convexity implies equivalence of goals.

Duality

- OMD can be analyzed using elementary tools.

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 $\|v\|$

Dual Vector Spaces

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

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

$$\|u\|_* \doteq \max_{v \in V} \frac{\|u(v)\|}{\|v\|}$$

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

$$\|u\|_* \doteq \max_{v \in V} \frac{\|u(v)\|}{\|v\|}$$

- V is equivalent to the set of all linear mappings from U to U.

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

$$\|u\|_* \doteq \max_{v \in V} \frac{\|u(v)\|}{\|v\|}$$

- 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 $U, V = \mathbb{R}^n$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- L_2 norm: $\sqrt{\sum_{i=1}^{n} x_i^2}$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- L_2 norm: $\sqrt{\sum_{i=1}^{n} x_i^2}$
- L_1 norm: $\sum_{i=1}^{n} |x_i|$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- L_2 norm: $\sqrt{\sum_{i=1}^n x_i^2}$
- L_1 norm: $\sum_{i=1}^n |x_i|$
- L_∞ norm: $\max_i |x_i|$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- 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: $\left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- 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 $\frac{1}{p} + \frac{1}{q} = 1$

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- 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.

Dual Norms

- The space is always $U, V = \mathbb{R}^n$
- The linear operation is the dot product $\mathrm{u} \cdot \mathrm{v}$
- $\|u\|_*$ is the dual norm to $\|v\|$,
- 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.

Lipschitz condition and the dual norm

Lemma 2.6:
Let $f : S \to \mathbb{R}$ be a convex function. Then, f is L-Lipschitz over S with respect to a norm $\|\cdot\|$ if and only if for all $w \in S$ and $z \in \partial f(w)$ we have:

$$\|z\|_* \leq L$$

where $\|\cdot\|_*$ denotes the dual norm.

Proof of Lemma 2.6

Proof:

Assume that f is L-Lipschitz. For any $w \in S$ and $z \in \partial f(w)$, choose u such that $u - w = \arg\max_{\|v\|=1} \langle v, z \rangle$. Then,

$$\langle z, u - w \rangle = \|z\|_*$$

By the sub-gradient definition,

$$f(u) - f(w) \geq \langle z, u - w \rangle = \|z\|_*$$

Since f is L-Lipschitz,

$$f(u) - f(w) \leq L\|u - w\| = L$$

Combining the inequalities:

$$\|z\|_* \leq L$$

For the converse, assume $\|z\|_* \leq L$ for all $z \in \partial f(w)$. Then,

$$f(u) - f(w) \leq \langle z, u - w \rangle \leq \|z\|_* \|u - w\| \leq L\|u - w\|$$

Hence, f is L-Lipschitz.

Fenchel Duality

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

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

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

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

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

- Fenchel duality Equivalent to Legendre duality for differentiable functions.

Visualization of the Legendre Dual

Visualization of the Legendre Dual

Visualization of the Legendre Dual

- $x, y\mathbb{R}$

Visualization of the Legendre Dual

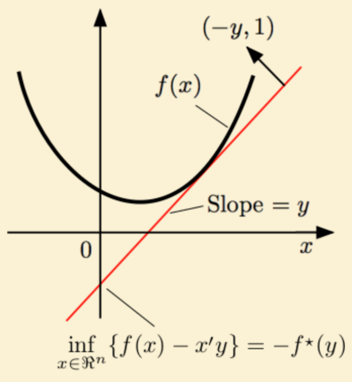
- $x, y\mathbb{R}$
- $f^*(y) = \sup_{x\in\mathbb{R}}(xy - f(x))$

Visualization of the Legendre Dual

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

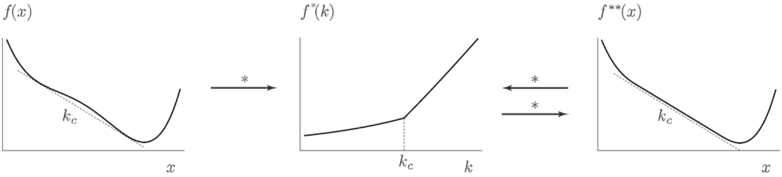
Visualization of the Legendre Dual

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



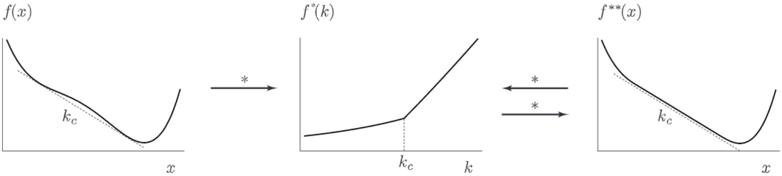
Dual of Dual

- The dual of any function is convex.



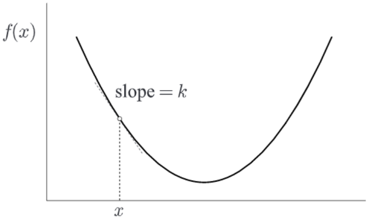
Dual of Dual

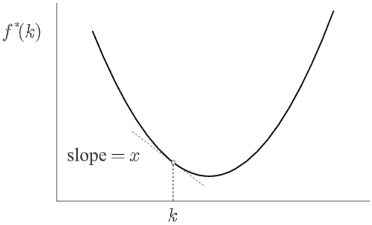
- The dual of any function is convex.
- if F is convex then $F^{**} = F$



Gradient Duality (legendre only)

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

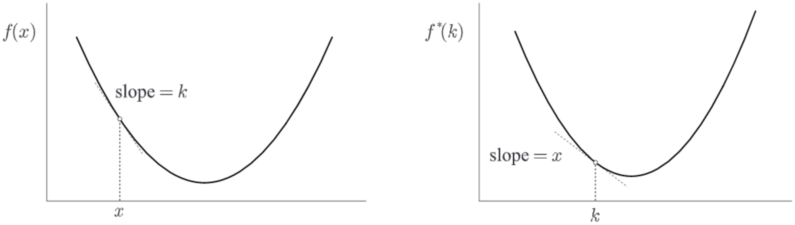




Gradient Duality (legendre only)

- 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: $F(\mathrm{u}) = \sum_{i=1}^{d} e^{u_i}$

Example: Exponential Potential

- Potential: $F(\mathrm{u}) = \sum_{i=1}^{d} e^{u_i}$
- Gradient: $\nabla F(\mathrm{u})_i = e^{u_i}$ or $\nabla F(\mathrm{u}) = F(\mathrm{u})$.

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: $F^*(\mathrm{v}) = \sum_{i=1}^{d} v_i(\ln v_i - 1)$

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: $\nabla F^*(\mathrm{v})_i = \ln v_i$

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

Bregman Divergence

Bregman Divergence

- $R(x)$ is convex and differentiable.

Bregman Divergence

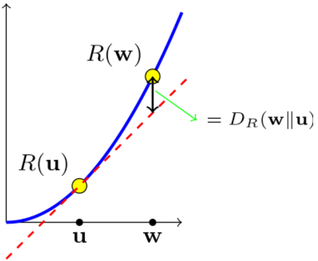
- $R(x)$ is convex and differentiable.
- $D_R(w||u) =$
  $R(w) - (R(u) + \langle \nabla R(u), (w - u) \rangle)$

Bregman Divergence

- $R(x)$ is convex and differentiable.
- $D_R(w||u) =$
  $R(w) - (R(u) + \langle \nabla R(u), (w - u) \rangle)$
- The error term of the first order Taylor expansion around u

Bregman Divergence

- $R(x)$ is convex and differentiable.
- $D_R(w||u) = R(w) - (R(u) + \langle \nabla R(u), (w-u) \rangle)$
- The error term of the first order Taylor expansion around u



Fenchel and Bregman

- F: strictly convex with continuous first derivative.

Fenchel and Bregman

- F: strictly convex with continuous first derivative.
- F^* is the Fenchel Dual of F

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^*

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 $\mathrm{v}' = \nabla F(\mathrm{v})$

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 - Step 1

Gradient Step in Dual Space:

$$z_{t+1} = \nabla R(w_t) - \eta \nabla f_t(w_t)$$

Here, $\nabla R(w_t)$ maps the point into the dual space.

Mirror Descent - Step 2

Projection Back to Primal Space:

$$w_{t+1} = \arg\min_{w \in S} D_R(w, z_{t+1})$$

Where $D_R(w, z)$ is the Bregman divergence:

$$D_R(w, z) = R(w) - R(z) - \langle \nabla R(z), w - z \rangle$$

This projection ensures w_{t+1} stays within the feasible set S.

Mirror Descent (alternative Notatio)

- Gradient descent in dual space $\theta_t = \theta_{t-1} - \lambda \nabla \ell_t(\theta_{t-1})$

Mirror Descent (alternative Notatio)

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

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

Mirror Descent (alternative Notatio)

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

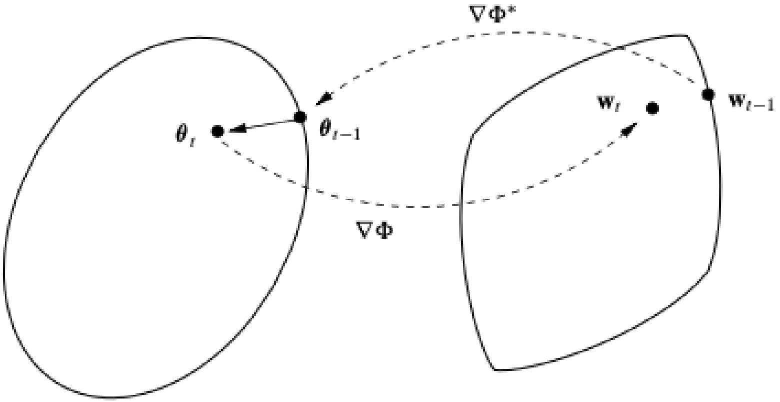
$$\nabla R^*(\mathrm{w}_t) = \nabla R^*(\mathrm{w}_{t-1}) - \lambda \nabla \ell_t(\mathrm{w}_{t-1})$$

- Projection: 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

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



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 $w_t = \nabla R^*(-z_{1:t-1})$ and using the definition of the Bregman divergence, we rewrite:

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

Rearranging, we get:

$$= -R^*(0) + \sum_{t=1}^{T} \left(\langle w_t, z_t \rangle - D_{R^*}(-z_{1:t} \| - z_{1:t-1})\right).$$

Final Step

Note: Since

$$R^*(0) = \max_{\mathrm{w}} \langle 0, \mathrm{w} \rangle - R(\mathrm{w}) = -\min_{\mathrm{w}} R(\mathrm{w}) = -R(\mathrm{w}_1),$$

combining all the above, we conclude the proof. $\square$

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$
- The dual norm for the weights is $\|w\|_2$

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$
- The dual norm for the weights is $\|w\|_2$
- We use the dual descend algorithm for the half quadratic regularizer

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$
- The dual norm for the weights is $\|w\|_2$
- We use the dual descend algorithm for the half quadratic regularizer
- Regularizer: $R(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$
- The dual norm for the weights is $\|w\|_2$
- We use the dual descend algorithm for the half quadratic regularizer
- Regularizer: $R(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$
- Legendre Dual Regularizer $R^*(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$
- The dual norm for the weights is $\|w\|_2$
- We use the dual descend algorithm for the half quadratic regularizer
- Regularizer: $R(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$
- Legendre Dual Regularizer $R^*(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$
- **Gradient Mapping:** $\nabla R(w) = w$

OMD for ℓ_2

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_2 \leq X_2$
- The dual norm for the weights is $\|w\|_2$
- We use the dual descend algorithm for the half quadratic regularizer
- Regularizer: $R(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$
- Legendre Dual Regularizer $R^*(\mathrm{u}) = \frac{1}{2}\|\mathrm{u}\|_2^2$
- **Gradient Mapping:** $\nabla R(w) = w$
- Gradient step: $z_{t+1} = w_t - \eta \nabla f_t(w_t)$

Projection Step for ℓ_2 Norm

Bregman Divergence:

$$D_R(w, z) = \frac{1}{2}\|w - z\|_2^2$$

Projection Back to Primal Space:

$$w_{t+1} = \Pi_S(z_{t+1}) = \arg\min_{w \in S} \frac{1}{2}\|w - z_{t+1}\|_2^2$$

Where Π_S denotes the Euclidean projection onto the feasible set S.

Final Update Rule for ℓ_2 Norm

Combining both steps, the final update rule becomes:

$$w_{t+1} = \Pi_S \left(w_t - \eta \nabla f_t(w_t) \right)$$

This is equivalent to the standard **Projected Gradient Descent** for the ℓ_2 norm.

Optimal Tuning for η and Regret Bound

Regret Bound:

$$\text{Regret}_T(u) \leq \frac{\|u\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(w_t)\|_2^2$$

Assuming $\|u\|_2 \leq B$ and $\|\nabla f_t(w_t)\|_2 \leq L$, this simplifies to:

$$\text{Regret}_T(u) \leq \frac{B^2}{2\eta} + \frac{\eta L^2 T}{2}$$

Optimal η:

$$\eta^* = \frac{B}{L\sqrt{T}}$$

Resulting Regret Bound:

$$\text{Regret}_T(u) \leq BL\sqrt{T}$$

ℓ_∞ OMD

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_\infty \leq X_\infty$

ℓ_∞ OMD

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_\infty \le X_\infty$
- The dual norm for the weights is $\|w\|_1$

ℓ_∞ OMD

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_\infty \leq X_\infty$
- The dual norm for the weights is $\|w\|_1$
- Suppose we use the dual descend algorithm for the exponential regularizer

ℓ_∞ OMD

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_\infty \leq X_\infty$
- The dual norm for the weights is $\|w\|_1$
- Suppose we use the dual descend algorithm for the exponential regularizer
- Regularizer: $R(\mathrm{u}) = \frac{1}{\eta} \ln \sum_{i=1}^{d} e^{\eta u_i}$

ℓ_∞ OMD

- Problem: OCO where $\forall z \in \partial f(x)\ \|z\|_\infty \le X_\infty$
- The dual norm for the weights is $\|w\|_1$
- Suppose we use the dual descend algorithm for the exponential regularizer
- Regularizer: $R(\mathrm{u}) = \frac{1}{\eta} \ln \sum_{i=1}^{d} e^{\eta u_i}$
- Legendre Dual Regularizer $R^*(\mathrm{u}) = \sum_{i=1}^{d} u_i(\ln u_i - 1)$

- Weight update:

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

- Weight update:

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

- Unnormalized exponentiated gradient:

$$w_{i,t} = w_{i,t-1} e^{-\lambda \nabla \ell_t (w_i - 1)}$$

- Weight update:

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

- Unnormalized exponentiated gradient:

$$w_{i,t} = w_{i,t-1} e^{-\lambda \nabla \ell_t (w_i - 1)}$$

- Normalized exponentiated gradient:

$$w_{i,t} = \frac{w_{i,t-1} e^{-\lambda \nabla \ell_t (w_i - 1)}}{\sum_{j=1}^{d} w_{j,t-1} e^{-\lambda \nabla \ell_t (w_j - 1)}}$$

- Normalization corresponds to projection on the simplex using the Bregman divergence according to R^*.

- Normalization corresponds to projection on the simplex using the Bregman divergence according to R^*.
- The dual descend algorithm for the exponential regularizer function R and the learning rate $\lambda = \frac{2\epsilon}{X_\infty^2}$ for some $0 < \epsilon < 1$

- Normalization corresponds to projection on the simplex using the Bregman divergence according to R^*.
- The dual descend algorithm for the exponential regularizer function R and the learning rate $\lambda = \frac{2\epsilon}{X_\infty^2}$ for some $0 < \epsilon < 1$
- yields Loss Bound:

$$L_{A,T} \leq \frac{L_T(\mathrm{u})}{1-\epsilon} + \frac{X_\infty^2 \ln d}{2\epsilon(1-\epsilon)}$$