Online Learning
Lecture 2

Nicolò Cesa-Bianchi

Università degli Studi di Milano

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable
- We can introduce curvature using a real-valued regularizer ψ
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- Example: the SVM objective function is $\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 + \frac{1}{m} \sum_{t=1}^{m} \ell_t(\boldsymbol{w})$
- If ℓ_t are all convex, this is equivalent to FTL over λ-strongly convex losses $\frac{\lambda}{2} \|\cdot\|_2^2 + \ell_t$
- How can we solve this constrained optimization problem?

FTRL with linearized losses

- Linearized regret: $R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) \leq \sum_{t=1}^{T} \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u})$

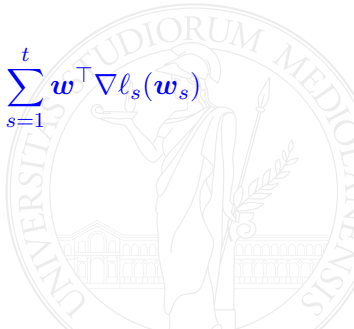
FTRL with linearized losses

- Linearized regret: $R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) \leq \sum_{t=1}^{T} \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u})$
- Pretend all losses are linear: $\ell'_t(\boldsymbol{w}) = \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$

FTRL with linearized losses

- Linearized regret: $R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) \leq \sum_{t=1}^{T} \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u})$
- Pretend all losses are linear: $\ell'_t(\boldsymbol{w}) = \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$
- FTRL with linearized losses:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell'_s(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \boldsymbol{w}^\top \nabla \ell_s(\boldsymbol{w}_s)$$

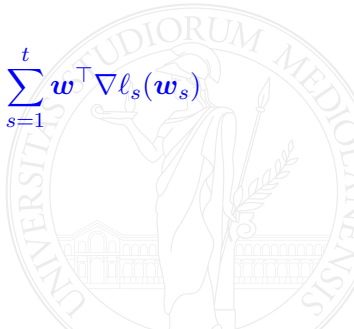


FTRL with linearized losses

- Linearized regret: $R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \Big) \leq \sum_{t=1}^{T} \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u})$
- Pretend all losses are linear: $\ell'_t(\boldsymbol{w}) = \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$
- FTRL with linearized losses:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell'_s(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \boldsymbol{w}^\top \nabla \ell_s(\boldsymbol{w}_s)$$

- We still have a constrained optimization problem to solve



Bregman divergences

- Assume ψ is convex

Bregman divergences

- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$

Bregman divergences

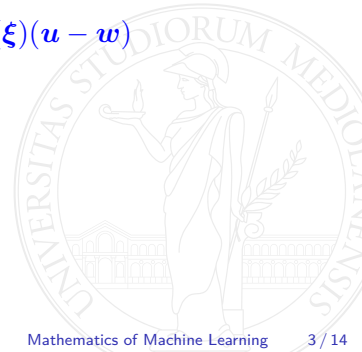
- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of ψ around $\boldsymbol{w}$

Bregman divergences

- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of ψ around $\boldsymbol{w}$
- If ψ is μ-strongly convex w.r.t. $\|\cdot\|$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$

Bregman divergences

- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of ψ around $\boldsymbol{w}$
- If ψ is μ-strongly convex w.r.t. $\|\cdot\|$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$
- If ψ is twice differentiable, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}(\boldsymbol{u} - \boldsymbol{w})^\top \nabla^2\psi(\boldsymbol{\xi})(\boldsymbol{u} - \boldsymbol{w})$ for some ξ on the line segment joining $\boldsymbol{u}$ and $\boldsymbol{w}$

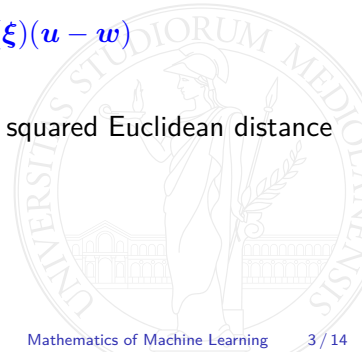


Bregman divergences

- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of ψ around $\boldsymbol{w}$
- If ψ is μ-strongly convex w.r.t. $\|\cdot\|$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$
- If ψ is twice differentiable, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}(\boldsymbol{u} - \boldsymbol{w})^\top \nabla^2\psi(\boldsymbol{\xi})(\boldsymbol{u} - \boldsymbol{w})$ for some ξ on the line segment joining $\boldsymbol{u}$ and $\boldsymbol{w}$
- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2$ squared Euclidean distance

Bregman divergences

- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of ψ around $\boldsymbol{w}$
- If ψ is μ-strongly convex w.r.t. $\|\cdot\|$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$
- If ψ is twice differentiable, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}(\boldsymbol{u} - \boldsymbol{w})^\top \nabla^2\psi(\boldsymbol{\xi})(\boldsymbol{u} - \boldsymbol{w})$ for some ξ on the line segment joining $\boldsymbol{u}$ and $\boldsymbol{w}$
- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2$ squared Euclidean distance
- Δ_d is the d-dimensional probability simplex



Bregman divergences

- Assume ψ is convex
- Bregman divergence: $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}) - \nabla\psi(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})$
- Error in first-order Taylor expansion of ψ around $\boldsymbol{w}$
- If ψ is μ-strongly convex w.r.t. $\|\cdot\|$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) \geq \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2$
- If ψ is twice differentiable, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}(\boldsymbol{u} - \boldsymbol{w})^\top \nabla^2\psi(\boldsymbol{\xi})(\boldsymbol{u} - \boldsymbol{w})$ for some ξ on the line segment joining $\boldsymbol{u}$ and $\boldsymbol{w}$
- If $\psi = \frac{1}{2}\|\cdot\|_2^2$, then $B_\psi(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{w}\|_2^2$ squared Euclidean distance
- Δ_d is the d-dimensional probability simplex
- If $\boldsymbol{p} \in \Delta_d$ and $\psi(\boldsymbol{p}) = \sum_i p(i) \ln p(i)$, then $B_\psi(\boldsymbol{p}, \boldsymbol{q}) = \sum_i p(i) \ln \frac{p(i)}{q(i)}$ KL-divergence

Bregman projections

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be strictly convex and differentiable in* $\mathbb{V}$.
Then $\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_f(\boldsymbol{w}, \boldsymbol{w}')$ *where* $\boldsymbol{w}' = \operatorname*{argmin}_{\boldsymbol{w}} f(\boldsymbol{w})$

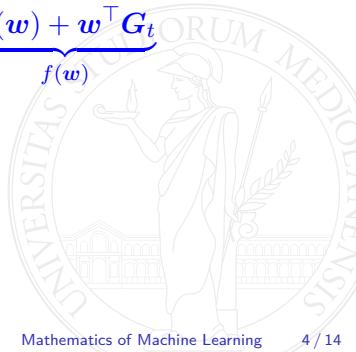
Bregman projections

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be strictly convex and differentiable in* $\mathbb{V}$.
Then $\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_f(\boldsymbol{w}, \boldsymbol{w}')$ *where* $\boldsymbol{w}' = \operatorname*{argmin}_{\boldsymbol{w}} f(\boldsymbol{w})$

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \boldsymbol{w}^\top \nabla \ell_s(\boldsymbol{w}_s) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \underbrace{\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t}_{f(\boldsymbol{w})}$$

where $\boldsymbol{G}_t = \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)$



Bregman projections

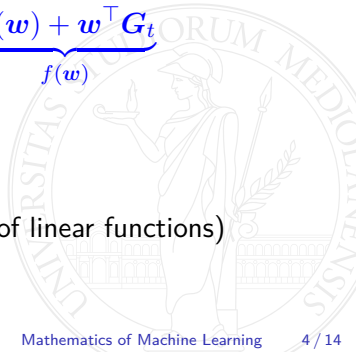
Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strictly convex and differentiable in $\mathbb{V}$.
Then $\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_f(\boldsymbol{w}, \boldsymbol{w}')$ where $\boldsymbol{w}' = \operatorname*{argmin}_{\boldsymbol{w}} f(\boldsymbol{w})$

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \boldsymbol{w}^\top \nabla \ell_s(\boldsymbol{w}_s) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \underbrace{\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t}_{f(\boldsymbol{w})}$$

where $\boldsymbol{G}_t = \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)$

Fact: $B_f \equiv B_\psi$ (because B_ψ is invariant with respect to addition of linear functions)



FTRL update in two steps:

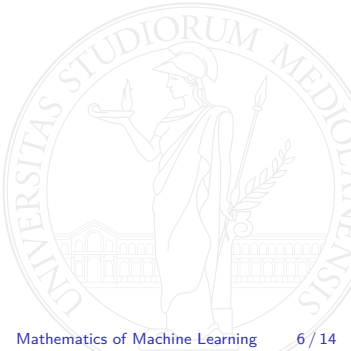
1. $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \boldsymbol{w}^\top \nabla \ell_s(\boldsymbol{w}_s)$

FTRL update in two steps:

1. $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \boldsymbol{w}^\top \nabla \ell_s(\boldsymbol{w}_s)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$ (Bregman projection)

Convex duality

- Assume ψ is convex



Convex duality

- Assume ψ is convex
- $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t \right) = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$ for $\boldsymbol{\theta}_{t+1} = -\boldsymbol{G}_t$

Convex duality

- Assume ψ is convex
- $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t \right) = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$ for $\boldsymbol{\theta}_{t+1} = -\boldsymbol{G}_t$
- The convex function $\psi^\star$ is the convex conjugate of ψ

$$\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta} - \psi(\boldsymbol{w}) \right)$$

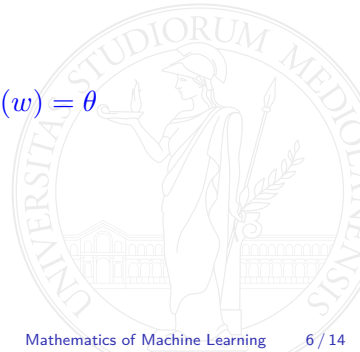
Convex duality

- Assume ψ is convex
- $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t \right) = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$ for $\boldsymbol{\theta}_{t+1} = -\boldsymbol{G}_t$
- The convex function $\psi^\star$ is the convex conjugate of ψ

$$\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta} - \psi(\boldsymbol{w}) \right)$$

- If ψ is differentiable

$$\frac{\partial}{\partial w} (w\theta - \psi(w)) = \theta - \psi'(w) = 0 \quad \text{iff} \quad \psi'(w) = \theta$$



Convex duality

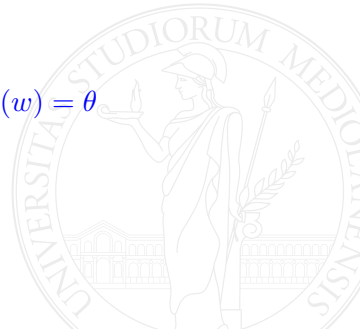
- Assume ψ is convex
- $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t \right) = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$ for $\boldsymbol{\theta}_{t+1} = -\boldsymbol{G}_t$
- The convex function $\psi^\star$ is the convex conjugate of ψ

$$\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta} - \psi(\boldsymbol{w}) \right)$$

- If ψ is differentiable

$$\frac{\partial}{\partial w}(w\theta - \psi(w)) = \theta - \psi'(w) = 0 \quad \text{iff} \quad \psi'(w) = \theta$$

- $w^* = \operatorname*{argmax}_{w \in \mathbb{R}} (w\theta - \psi(w))$



Convex duality

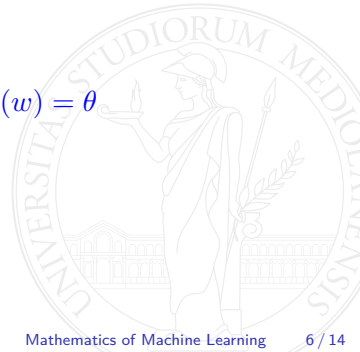
- Assume ψ is convex
- $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{G}_t \right) = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$ for $\boldsymbol{\theta}_{t+1} = -\boldsymbol{G}_t$
- The convex function $\psi^\star$ is the convex conjugate of ψ

$$\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta} - \psi(\boldsymbol{w}) \right)$$

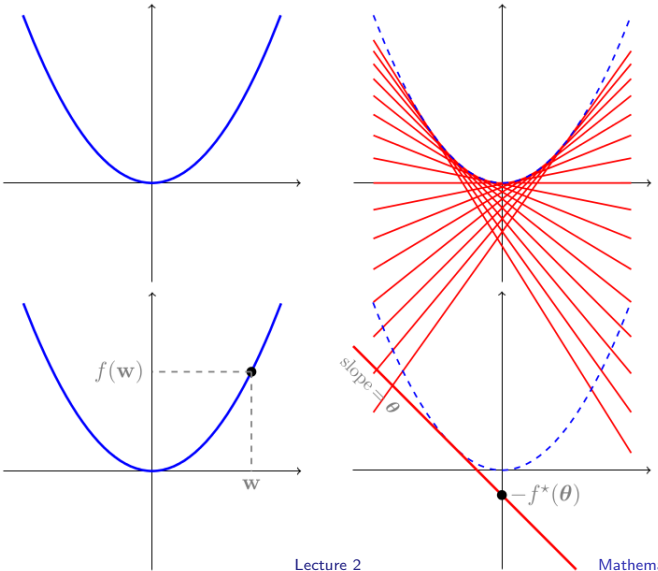
- If ψ is differentiable

$$\frac{\partial}{\partial w}(w\theta - \psi(w)) = \theta - \psi'(w) = 0 \quad \text{iff} \quad \psi'(w) = \theta$$

- $w^* = \operatorname*{argmax}_{w \in \mathbb{R}} (w\theta - \psi(w))$
- $\psi(w^*) = w^*\theta - \psi^\star(\theta)$



Convex duality



Strongly convex regularizers

- $\boldsymbol{w}'_{t+1} = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$

Strongly convex regularizers

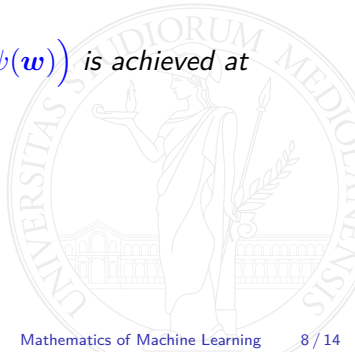
- $\boldsymbol{w}'_{t+1} = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$
- $\psi^\star(\boldsymbol{\theta}_{t+1}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$

Strongly convex regularizers

- $\boldsymbol{w}'_{t+1} = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$
- $\psi^\star(\boldsymbol{\theta}_{t+1}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$

Theorem

If ψ is strongly convex, then $\psi^\star$ is differentiable and $\max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta} - \psi(\boldsymbol{w}) \right)$ is achieved at $\boldsymbol{w}^* = \nabla \psi^\star(\boldsymbol{\theta})$



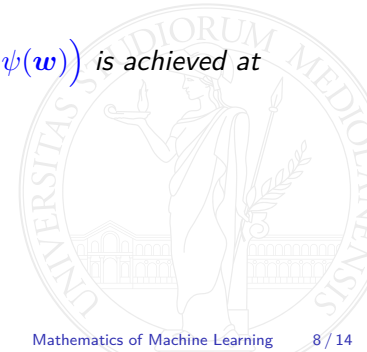
Strongly convex regularizers

- $\boldsymbol{w}'_{t+1} = \operatorname*{argmax}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$
- $\psi^\star(\boldsymbol{\theta}_{t+1}) = \max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w}) \right)$

Theorem

If ψ is strongly convex, then $\psi^\star$ is differentiable and $\max_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\theta} - \psi(\boldsymbol{w}) \right)$ is achieved at $\boldsymbol{w}^* = \nabla \psi^\star(\boldsymbol{\theta})$

This implies $\boldsymbol{w}'_{t+1} = \nabla \psi^\star(\boldsymbol{\theta}_{t+1})$ for ψ strongly convex



FTRL update (for strongly convex regularizers)

1. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \nabla \ell_t(\boldsymbol{w}_t)$ (gradient update)

FTRL update (for strongly convex regularizers)

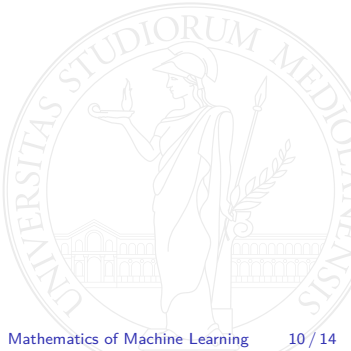
1. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \nabla \ell_t(\boldsymbol{w}_t)$ (gradient update)
2. $\boldsymbol{w}'_{t+1} = \nabla \psi^\star(\boldsymbol{\theta}_{t+1})$ (mirror mapping)

FTRL update (for strongly convex regularizers)

1. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \nabla \ell_t(\boldsymbol{w}_t)$ (gradient update)
2. $\boldsymbol{w}'_{t+1} = \nabla \psi^\star(\boldsymbol{\theta}_{t+1})$ (mirror mapping)
3. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$ (Bregman projection)

The Euclidean case: Lazy Online Gradient Descent

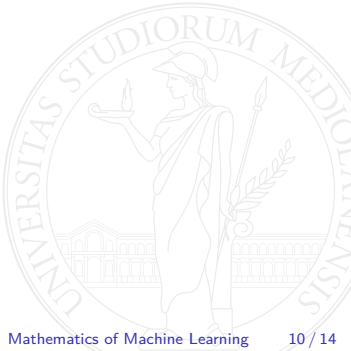
- $\psi = \frac{1}{2\eta} \|\cdot\|_2^2$ strongly convex w.r.t. $\|\cdot\|_2$ ($\eta > 0$ is a learning rate)



The Euclidean case: Lazy Online Gradient Descent

- $\psi = \frac{1}{2\eta} \|\cdot\|_2^2$
- $\psi^\star = \frac{\eta}{2} \|\cdot\|_2^2$

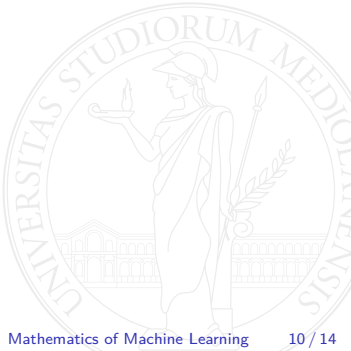
strongly convex w.r.t. $\|\cdot\|_2$ ($\eta > 0$ is a learning rate)



The Euclidean case: Lazy Online Gradient Descent

- $\psi = \frac{1}{2\eta} \|\cdot\|_2^2$
- $\psi^\star = \frac{\eta}{2} \|\cdot\|_2^2$
- $\nabla \psi^\star(\boldsymbol{\theta}) = \eta \boldsymbol{\theta}$

strongly convex w.r.t. $\|\cdot\|_2$ ($\eta > 0$ is a learning rate)



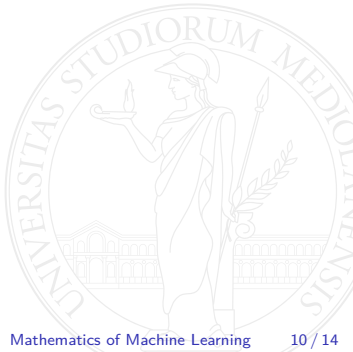
The Euclidean case: Lazy Online Gradient Descent

- $\psi = \frac{1}{2\eta} \|\cdot\|_2^2$
- $\psi^\star = \frac{\eta}{2} \|\cdot\|_2^2$
- $\nabla \psi^\star(\boldsymbol{\theta}) = \eta \boldsymbol{\theta}$

strongly convex w.r.t. $\|\cdot\|_2$ ($\eta > 0$ is a learning rate)

FTRL update (Projected Lazy OGD):

1. $\boldsymbol{w}'_{t+1} = -\eta \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)$

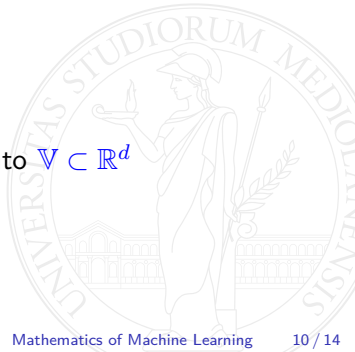


The Euclidean case: Lazy Online Gradient Descent

- $\psi = \frac{1}{2\eta} \|\cdot\|_2^2$ strongly convex w.r.t. $\|\cdot\|_2$ ($\eta > 0$ is a learning rate)
- $\psi^\star = \frac{\eta}{2} \|\cdot\|_2^2$
- $\nabla \psi^\star(\boldsymbol{\theta}) = \eta \boldsymbol{\theta}$

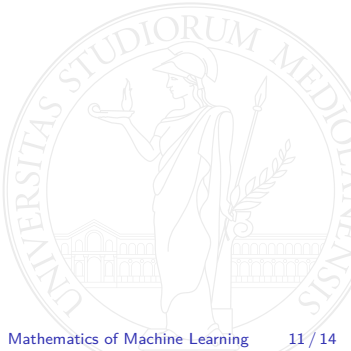
FTRL update (Projected Lazy OGD):

1. $\boldsymbol{w}'_{t+1} = -\eta \sum_{s=1}^{t} \nabla \ell_s(\boldsymbol{w}_s)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \left\| \boldsymbol{w} - \boldsymbol{w}'_{t+1} \right\|_2$ standard Euclidean projection onto $\mathbb{V} \subset \mathbb{R}^d$



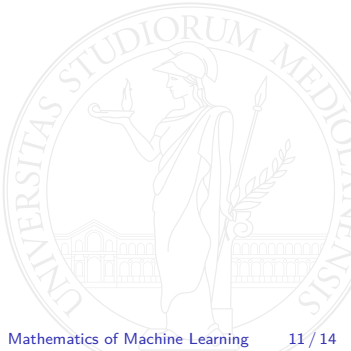
The Entropic case: Exponentiated Gradient

- $\mathbb{V} = \Delta_d$ (probability simplex)



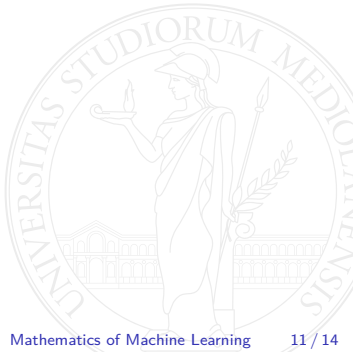
The Entropic case: Exponentiated Gradient

- $\mathbb{V} = \Delta_d$ (probability simplex)
- $\psi(\boldsymbol{p}) = \frac{1}{\eta} \sum_i p(i) \ln p(i)$ for $\boldsymbol{p} \in \Delta_d$ strongly convex w.r.t. $\|\cdot\|_1$



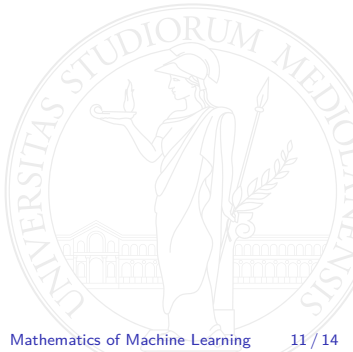
The Entropic case: Exponentiated Gradient

- $\mathbb{V} = \Delta_d$ (probability simplex)
- $\psi(\boldsymbol{p}) = \frac{1}{\eta} \sum_i p(i) \ln p(i)$ for $\boldsymbol{p} \in \Delta_d$ strongly convex w.r.t. $\|\cdot\|_1$
- $\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{p} \in \Delta_d} \left(\boldsymbol{p}^\top \boldsymbol{\theta} - \psi(\boldsymbol{p}) \right) = \frac{1}{\eta} \ln \left(\sum_i e^{\eta \theta(i)} \right)$ we solve the constrained problem



The Entropic case: Exponentiated Gradient

- $\mathbb{V} = \Delta_d$ (probability simplex)
- $\psi(\boldsymbol{p}) = \frac{1}{\eta} \sum_i p(i) \ln p(i)$ for $\boldsymbol{p} \in \Delta_d$ strongly convex w.r.t. $\|\cdot\|_1$
- $\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{p} \in \Delta_d} \left(\boldsymbol{p}^\top \boldsymbol{\theta} - \psi(\boldsymbol{p}) \right) = \frac{1}{\eta} \ln \left(\sum_i e^{\eta \theta(i)} \right)$ we solve the constrained problem
- $\nabla \psi^\star(\boldsymbol{\theta})_i = \frac{e^{\eta \theta(i)}}{\sum_{j=1}^d e^{\eta \theta(j)}}$

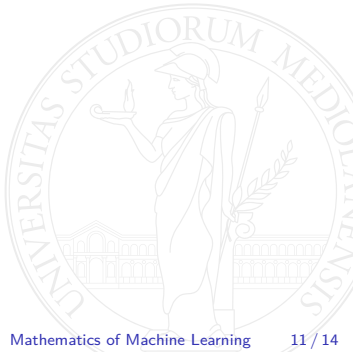


The Entropic case: Exponentiated Gradient

- $\mathbb{V} = \Delta_d$ (probability simplex)
- $\psi(\boldsymbol{p}) = \frac{1}{\eta} \sum_i p(i) \ln p(i)$ for $\boldsymbol{p} \in \Delta_d$ strongly convex w.r.t. $\|\cdot\|_1$
- $\psi^\star(\boldsymbol{\theta}) = \max_{\boldsymbol{p} \in \Delta_d} \left(\boldsymbol{p}^\top \boldsymbol{\theta} - \psi(\boldsymbol{p}) \right) = \frac{1}{\eta} \ln \left(\sum_i e^{\eta \theta(i)} \right)$ we solve the constrained problem
- $\nabla \psi^\star(\boldsymbol{\theta})_i = \dfrac{e^{\eta \theta(i)}}{\sum_{j=1}^d e^{\eta \theta(j)}}$

FTRL update (EG):

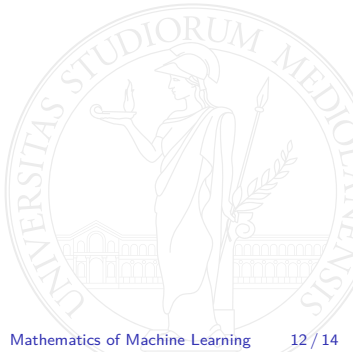
$$p_{t+1}(i) = \frac{\exp\left(-\eta \sum_{s=1}^t \nabla \ell_s(\boldsymbol{p}_s)_i\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^t \nabla \ell_s(\boldsymbol{p}_s)_j\right)}$$



Online Mirror Descent

Update minimizes trade-off between linearized loss and Bregman divergence from previous iterate

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \, \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$$



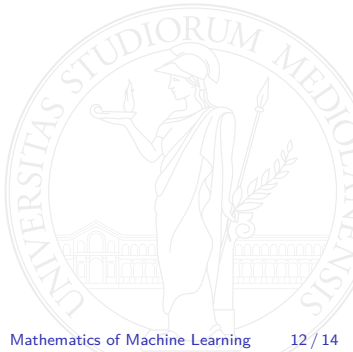
Online Mirror Descent

Update minimizes trade-off between linearized loss and Bregman divergence from previous iterate

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \, \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$$

Projected update:

1. $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \, \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$



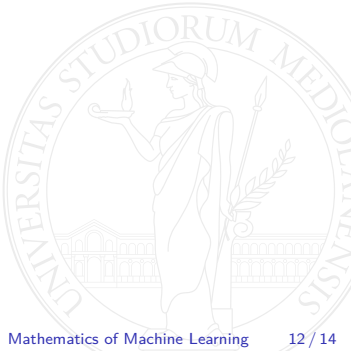
Online Mirror Descent

Update minimizes trade-off between linearized loss and Bregman divergence from previous iterate

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \, \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$$

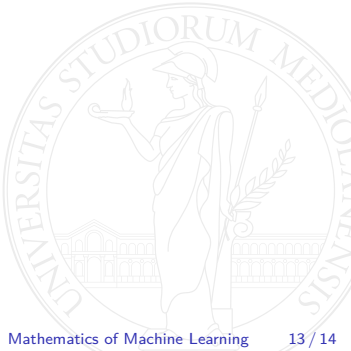
Projected update:

1. $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \, \boldsymbol{w}^\top \nabla \ell_t(\boldsymbol{w}_t)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$



OMD: computation of the update

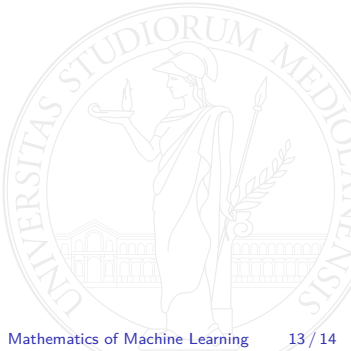
Assume ψ is strongly convex



OMD: computation of the update

Assume ψ is strongly convex

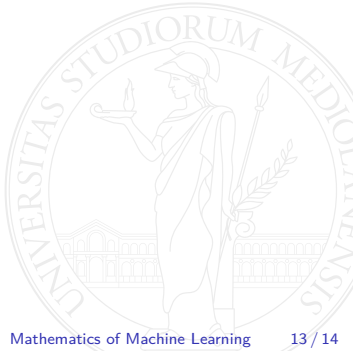
$$\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \qquad (\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t))$$



OMD: computation of the update

Assume ψ is strongly convex

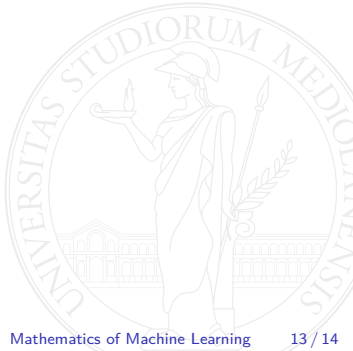
$$\begin{aligned} \boldsymbol{w}'_{t+1} &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \qquad (\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)) \\ &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) - \psi(\boldsymbol{w}_t) - \nabla\psi(\boldsymbol{w}_t)^\top(\boldsymbol{w} - \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \end{aligned}$$



OMD: computation of the update

Assume ψ is strongly convex

$$
\begin{aligned}
\boldsymbol{w}'_{t+1} &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t && (\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)) \\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) - \psi(\boldsymbol{w}_t) - \nabla\psi(\boldsymbol{w}_t)^\top (\boldsymbol{w} - \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) + \boldsymbol{w}^\top \Big(\eta_t \, \boldsymbol{g}_t - \nabla\psi(\boldsymbol{w}_t)\Big)
\end{aligned}
$$

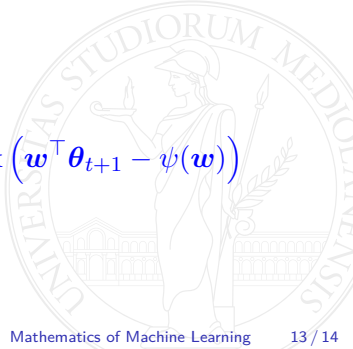


OMD: computation of the update

Assume ψ is strongly convex

$$
\begin{aligned}
\boldsymbol{w}'_{t+1} &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \qquad (\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)) \\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) - \psi(\boldsymbol{w}_t) - \nabla\psi(\boldsymbol{w}_t)^\top(\boldsymbol{w} - \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) + \boldsymbol{w}^\top \Big(\eta_t \boldsymbol{g}_t - \nabla\psi(\boldsymbol{w}_t)\Big)
\end{aligned}
$$

We use $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \Big(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \Big(\eta_t \boldsymbol{g}_t - \nabla\psi(\boldsymbol{w}_t)\Big)\Big) = \operatorname*{argmax}_{\boldsymbol{w}} \Big(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w})\Big)$
for $\boldsymbol{\theta}_{t+1} = \nabla\psi(\boldsymbol{w}_t) - \eta_t \boldsymbol{g}_t$



OMD: computation of the update

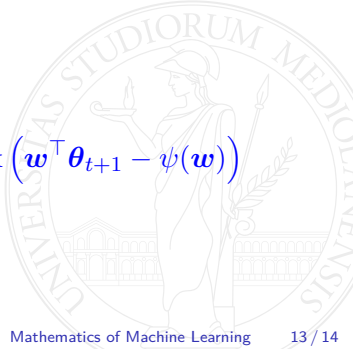
Assume ψ is strongly convex

$$\begin{aligned}
\boldsymbol{w}'_{t+1} &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t && (\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)) \\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) - \psi(\boldsymbol{w}_t) - \nabla\psi(\boldsymbol{w}_t)^\top(\boldsymbol{w} - \boldsymbol{w}_t) + \eta_t \boldsymbol{w}^\top \boldsymbol{g}_t \\
&= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \psi(\boldsymbol{w}) + \boldsymbol{w}^\top \Big(\eta_t \boldsymbol{g}_t - \nabla\psi(\boldsymbol{w}_t)\Big)
\end{aligned}$$

We use $\boldsymbol{w}'_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \Big(\psi(\boldsymbol{w}) + \boldsymbol{w}^\top \Big(\eta_t \boldsymbol{g}_t - \nabla\psi(\boldsymbol{w}_t)\Big)\Big) = \operatorname*{argmax}_{\boldsymbol{w}} \Big(\boldsymbol{w}^\top \boldsymbol{\theta}_{t+1} - \psi(\boldsymbol{w})\Big)$

for $\boldsymbol{\theta}_{t+1} = \nabla\psi(\boldsymbol{w}_t) - \eta_t \boldsymbol{g}_t$

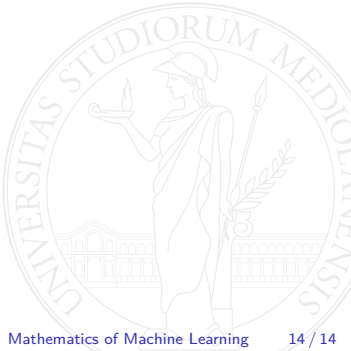
$\boldsymbol{w}'_{t+1} = \nabla\psi^\star(\boldsymbol{\theta}_{t+1}) = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t \boldsymbol{g}_t\Big)$



Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla \psi^\star \left(- \sum_{s=1}^{t} \boldsymbol{g}_s \right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

1. $\boldsymbol{w}'_{t+1} = \nabla \psi^\star \Big(\nabla \psi(\boldsymbol{w}_t) - \eta_t \, \boldsymbol{g}_t \Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

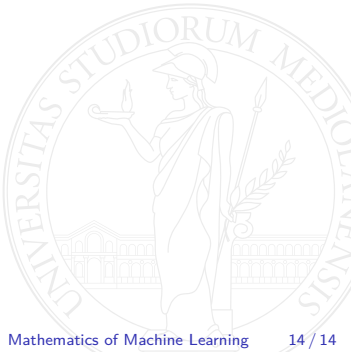


Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\left(-\sum_{s=1}^{t}\boldsymbol{g}_s\right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

- $\nabla\psi^\star$ maps gradients to iterates $\boldsymbol{w}_t$

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t\,\boldsymbol{g}_t\Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

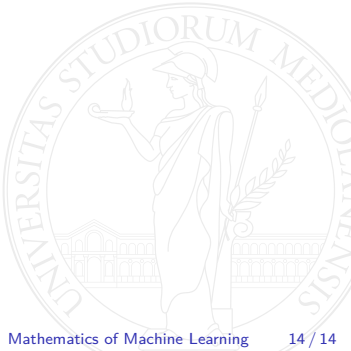


Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\left(-\sum_{s=1}^{t} \boldsymbol{g}_s\right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

- $\nabla\psi^\star$ maps gradients to iterates $\boldsymbol{w}_t$
- $\nabla\psi$ maps iterates to gradients

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t\, \boldsymbol{g}_t\Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

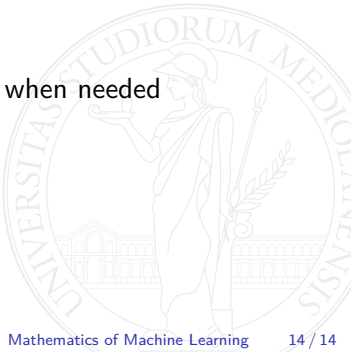


Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\left(-\sum_{s=1}^{t} \boldsymbol{g}_s\right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t\, \boldsymbol{g}_t\Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

- $\nabla\psi^\star$ maps gradients to iterates $\boldsymbol{w}_t$
- $\nabla\psi$ maps iterates to gradients
- FTRL updates a state variable $-\sum_{s=1}^{t} \boldsymbol{g}_s$ and maps it to iterates when needed

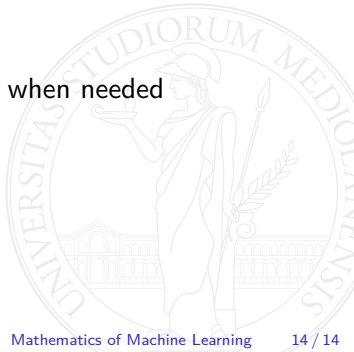


Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\left(-\sum_{s=1}^{t}\boldsymbol{g}_s\right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t\,\boldsymbol{g}_t\Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

- $\nabla\psi^\star$ maps gradients to iterates $\boldsymbol{w}_t$
- $\nabla\psi$ maps iterates to gradients
- FTRL updates a state variable $-\sum_{s=1}^{t}\boldsymbol{g}_s$ and maps it to iterates when needed
- OMD maps iterates back to gradients before each update

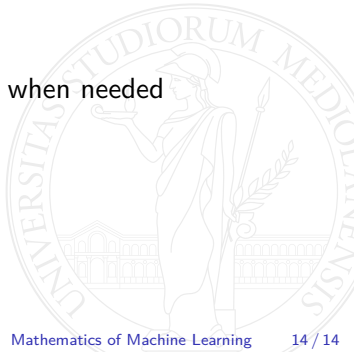


Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\left(-\sum_{s=1}^{t}\boldsymbol{g}_s\right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t\,\boldsymbol{g}_t\Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

- $\nabla\psi^\star$ maps gradients to iterates $\boldsymbol{w}_t$
- $\nabla\psi$ maps iterates to gradients
- FTRL updates a state variable $-\sum_{s=1}^{t}\boldsymbol{g}_s$ and maps it to iterates when needed
- OMD maps iterates back to gradients before each update
- OMD and FTRL have similar regret bounds in many cases



Some differences between FTRL and OMD

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\left(-\sum_{s=1}^{t}\boldsymbol{g}_s\right)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

1. $\boldsymbol{w}'_{t+1} = \nabla\psi^\star\Big(\nabla\psi(\boldsymbol{w}_t) - \eta_t\,\boldsymbol{g}_t\Big)$
2. $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} B_\psi(\boldsymbol{w}, \boldsymbol{w}'_{t+1})$

- $\nabla\psi^\star$ maps gradients to iterates $\boldsymbol{w}_t$
- $\nabla\psi$ maps iterates to gradients
- FTRL updates a state variable $-\sum_{s=1}^{t}\boldsymbol{g}_s$ and maps it to iterates when needed
- OMD maps iterates back to gradients before each update
- OMD and FTRL have similar regret bounds in many cases
- In certain cases, FTRL works better than OMD when using dynamic learning rates