

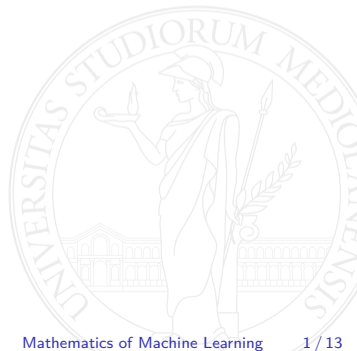
Online Learning

Lecture 5

Nicolò Cesa-Bianchi

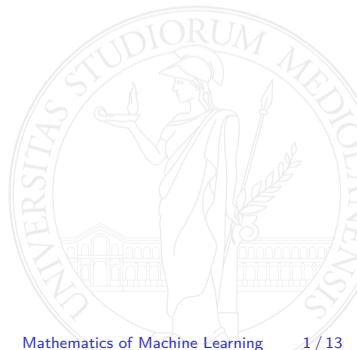
Università degli Studi di Milano

Exploiting curvature of the losses



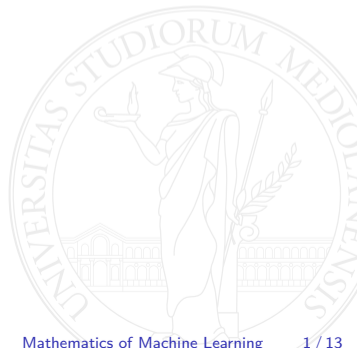
Exploiting curvature of the losses

- ▶ Convex and G -Lipschitz losses: FTRL with $\psi = \frac{1}{2} \|\cdot\|_2^2$ achieves $R_T = \mathcal{O}(GD\sqrt{T})$



Exploiting curvature of the losses

- ▶ Convex and G -Lipschitz losses: FTRL with $\psi = \frac{1}{2} \|\cdot\|_2^2$ achieves $R_T = \mathcal{O}(GD\sqrt{T})$
- ▶ Strongly convex and G -Lipschitz losses: FTL achieves $R_T = \mathcal{O}(G^2 \ln T)$



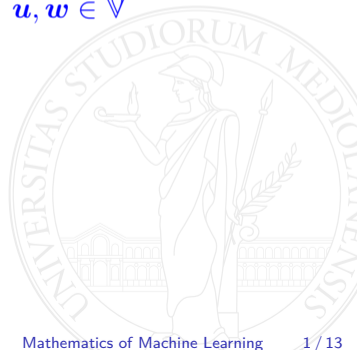
Exploiting curvature of the losses

- ▶ Convex and G -Lipschitz losses: FTRL with $\psi = \frac{1}{2} \|\cdot\|_2^2$ achieves $R_T = \mathcal{O}(GD\sqrt{T})$
- ▶ Strongly convex and G -Lipschitz losses: FTL achieves $R_T = \mathcal{O}(G^2 \ln T)$

Strong convexity in the direction of the gradient (exp-concavity)

$$\ell_t(\mathbf{u}) \geq \ell_t(\mathbf{w}) + \mathbf{g}_t^\top (\mathbf{u} - \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{w}\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{u}, \mathbf{w} \in \mathbb{V}$$

where $\mathbf{g}_t = \nabla \ell_t(\mathbf{w})$ and $\|\mathbf{w}\|_M^2 = \mathbf{w}^\top M \mathbf{w}$



Exploiting curvature of the losses

- ▶ Convex and G -Lipschitz losses: FTRL with $\psi = \frac{1}{2} \|\cdot\|_2^2$ achieves $R_T = \mathcal{O}(GD\sqrt{T})$
- ▶ Strongly convex and G -Lipschitz losses: FTL achieves $R_T = \mathcal{O}(G^2 \ln T)$

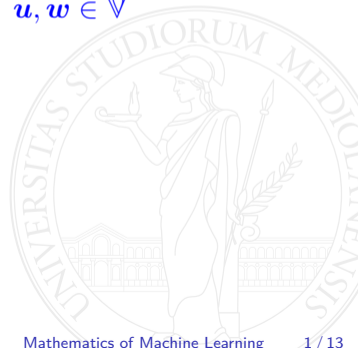
Strong convexity in the direction of the gradient (exp-concavity)

$$\ell_t(\mathbf{u}) \geq \ell_t(\mathbf{w}) + \mathbf{g}_t^\top (\mathbf{u} - \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{w}\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{u}, \mathbf{w} \in \mathbb{V}$$

where $\mathbf{g}_t = \nabla \ell_t(\mathbf{w})$ and $\|\mathbf{w}\|_M^2 = \mathbf{w}^\top M \mathbf{w}$

Some losses satisfying the condition (in a bounded domain)

- ▶ Square loss $\ell(\mathbf{w}) = \frac{1}{2} (\mathbf{w}^\top \mathbf{x} - y)^2$



Exploiting curvature of the losses

- ▶ Convex and G -Lipschitz losses: FTRL with $\psi = \frac{1}{2} \|\cdot\|_2^2$ achieves $R_T = \mathcal{O}(GD\sqrt{T})$
- ▶ Strongly convex and G -Lipschitz losses: FTL achieves $R_T = \mathcal{O}(G^2 \ln T)$

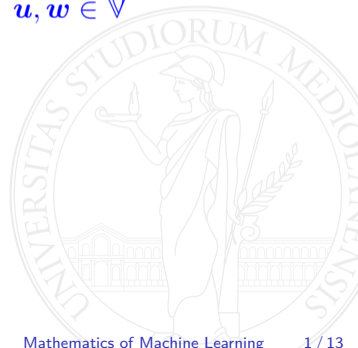
Strong convexity in the direction of the gradient (exp-concavity)

$$\ell_t(\mathbf{u}) \geq \ell_t(\mathbf{w}) + \mathbf{g}_t^\top (\mathbf{u} - \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{w}\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{u}, \mathbf{w} \in \mathbb{V}$$

where $\mathbf{g}_t = \nabla \ell_t(\mathbf{w})$ and $\|\mathbf{w}\|_M^2 = \mathbf{w}^\top M \mathbf{w}$

Some losses satisfying the condition (in a bounded domain)

- ▶ Square loss $\ell(\mathbf{w}) = \frac{1}{2} (\mathbf{w}^\top \mathbf{x} - y)^2$
- ▶ Logistic loss $\ell(\mathbf{w}) = \ln(1 + \exp(-y \mathbf{w}^\top \mathbf{x}))$



Online Newton Step for exp-concave losses

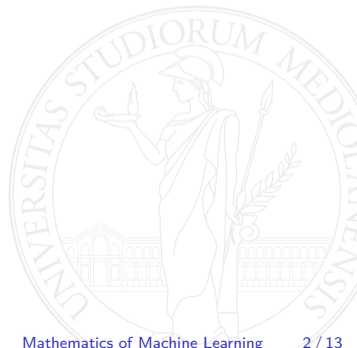
Choose the model minimizing a second-order approximation of the true loss:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{s=1}^t \hat{\ell}_s(\mathbf{w})$$

(FTL on a surrogate loss)

$$\hat{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top (\mathbf{w} - \mathbf{w}_t) + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_t\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2$$

$$\mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$$



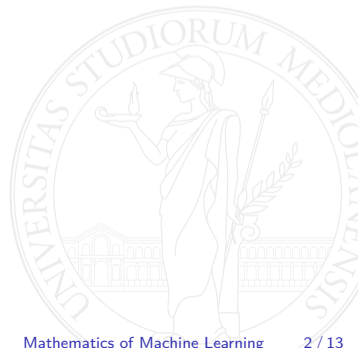
Online Newton Step for exp-concave losses

Choose the model minimizing a second-order approximation of the true loss:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{s=1}^t \hat{\ell}_s(\mathbf{w}) \quad (\text{FTL on a surrogate loss})$$

$$\hat{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top (\mathbf{w} - \mathbf{w}_t) + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_t\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$$

Properties:



Online Newton Step for exp-concave losses

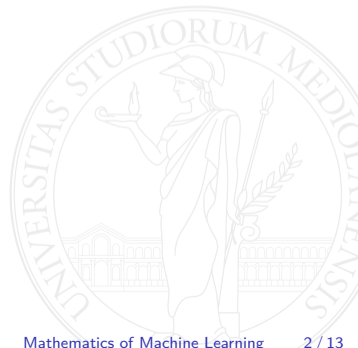
Choose the model minimizing a second-order approximation of the true loss:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{s=1}^t \hat{\ell}_s(\mathbf{w}) \quad (\text{FTL on a surrogate loss})$$

$$\hat{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top (\mathbf{w} - \mathbf{w}_t) + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_t\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$$

Properties:

- ▶ $\hat{\ell}_t(\mathbf{u}) \leq \ell_t(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{V}$



Online Newton Step for exp-concave losses

Choose the model minimizing a second-order approximation of the true loss:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{s=1}^t \widehat{\ell}_s(\mathbf{w})$$

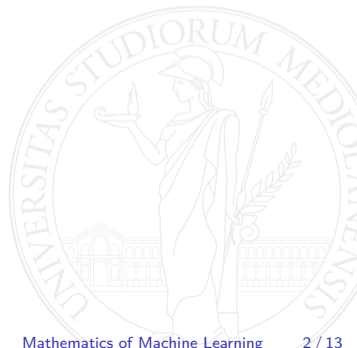
(FTL on a surrogate loss)

$$\widehat{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top (\mathbf{w} - \mathbf{w}_t) + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_t\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2$$

$$\mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$$

Properties:

- ▶ $\widehat{\ell}_t(\mathbf{u}) \leq \ell_t(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{V}$
- ▶ $\widehat{\ell}_t(\mathbf{w}_t) = \ell_t(\mathbf{w}_t)$



Online Newton Step for exp-concave losses

Choose the model minimizing a second-order approximation of the true loss:

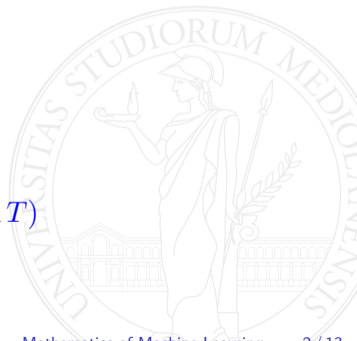
$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{s=1}^t \hat{\ell}_s(\mathbf{w}) \quad (\text{FTL on a surrogate loss})$$

$$\hat{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top (\mathbf{w} - \mathbf{w}_t) + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_t\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$$

Properties:

- ▶ $\hat{\ell}_t(\mathbf{u}) \leq \ell_t(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{V}$
- ▶ $\hat{\ell}_t(\mathbf{w}_t) = \ell_t(\mathbf{w}_t)$

- ▶ Regret bound:
$$R_T(\mathbf{u}) \leq \sum_{t=1}^T \hat{\ell}_t(\mathbf{w}_t) - \sum_{t=1}^T \hat{\ell}_t(\mathbf{u}) = \mathcal{O}(GDd \ln T)$$



Online Newton Step for exp-concave losses

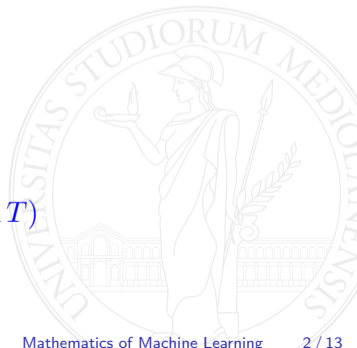
Choose the model minimizing a second-order approximation of the true loss:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{V}} \sum_{s=1}^t \hat{\ell}_s(\mathbf{w}) \quad (\text{FTL on a surrogate loss})$$

$$\hat{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \mathbf{g}_t^\top (\mathbf{w} - \mathbf{w}_t) + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_t\|_{\mathbf{g}_t \mathbf{g}_t^\top}^2 \quad \mathbf{g}_t = \nabla \ell_t(\mathbf{w}_t)$$

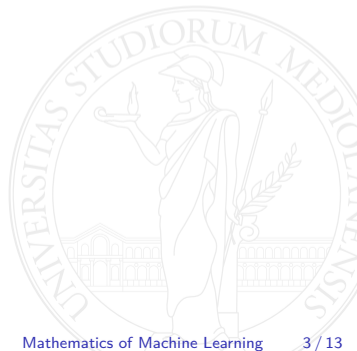
Properties:

- ▶ $\hat{\ell}_t(\mathbf{u}) \leq \ell_t(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{V}$
- ▶ $\hat{\ell}_t(\mathbf{w}_t) = \ell_t(\mathbf{w}_t)$
- ▶ Regret bound: $R_T(\mathbf{u}) \leq \sum_{t=1}^T \hat{\ell}_t(\mathbf{w}_t) - \sum_{t=1}^T \hat{\ell}_t(\mathbf{u}) = \mathcal{O}(GDd \ln T)$
- ▶ This matches the $\mathcal{O}(\ln T)$ bound for strongly convex losses



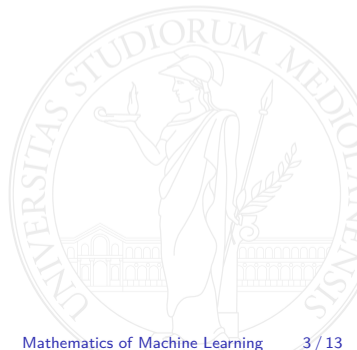
Unconstrained online convex optimization

- ▶ Assume l_t is 1-Lipschitz for $t \geq 1$



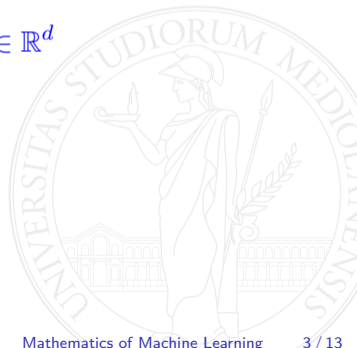
Unconstrained online convex optimization

- ▶ Assume l_t is 1-Lipschitz for $t \geq 1$
- ▶ Run FTRL with Euclidean regularizer $\psi = \frac{1}{2} \|\cdot\|_2^2$, no projection, and learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$



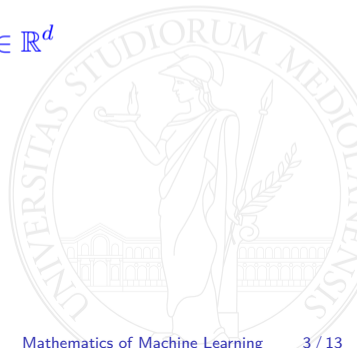
Unconstrained online convex optimization

- ▶ Assume l_t is 1-Lipschitz for $t \geq 1$
- ▶ Run FTRL with Euclidean regularizer $\psi = \frac{1}{2} \|\cdot\|_2^2$, no projection, and learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- ▶ $R_T(\mathbf{u}) \leq \frac{\psi(\mathbf{u}) - \psi(\mathbf{w}_1)}{\eta} + \eta T = \frac{1}{2} \left(\frac{\|\mathbf{u}\|_2^2}{\alpha} + \alpha \right) \sqrt{T} \quad \forall \mathbf{u} \in \mathbb{R}^d$



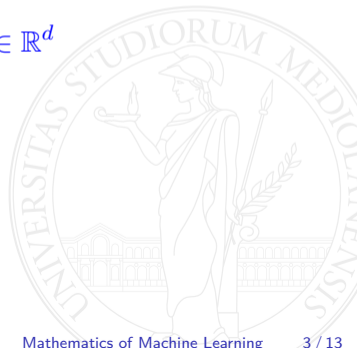
Unconstrained online convex optimization

- ▶ Assume l_t is 1-Lipschitz for $t \geq 1$
- ▶ Run FTRL with Euclidean regularizer $\psi = \frac{1}{2} \|\cdot\|_2^2$, no projection, and learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- ▶ $R_T(\mathbf{u}) \leq \frac{\psi(\mathbf{u}) - \psi(\mathbf{w}_1)}{\eta} + \eta T = \frac{1}{2} \left(\frac{\|\mathbf{u}\|_2^2}{\alpha} + \alpha \right) \sqrt{T} \quad \forall \mathbf{u} \in \mathbb{R}^d$
- ▶ $R_T(\mathbf{u}) \leq \|\mathbf{u}\|_2 \sqrt{T}$ for $\alpha = \|\mathbf{u}\|_2$



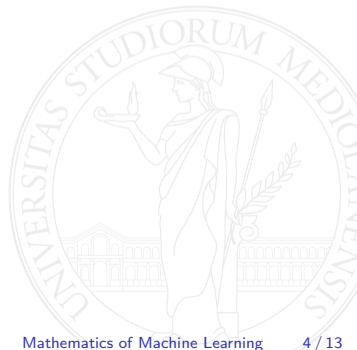
Unconstrained online convex optimization

- ▶ Assume l_t is 1-Lipschitz for $t \geq 1$
- ▶ Run FTRL with Euclidean regularizer $\psi = \frac{1}{2} \|\cdot\|_2^2$, no projection, and learning rate $\eta = \alpha/\sqrt{T}$ for $\alpha > 0$
- ▶ $R_T(\mathbf{u}) \leq \frac{\psi(\mathbf{u}) - \psi(\mathbf{w}_1)}{\eta} + \eta T = \frac{1}{2} \left(\frac{\|\mathbf{u}\|_2^2}{\alpha} + \alpha \right) \sqrt{T} \quad \forall \mathbf{u} \in \mathbb{R}^d$
- ▶ $R_T(\mathbf{u}) \leq \|\mathbf{u}\|_2 \sqrt{T}$ for $\alpha = \|\mathbf{u}\|_2$
- ▶ This bound cannot be simultaneously achieved for all \mathbf{u} !



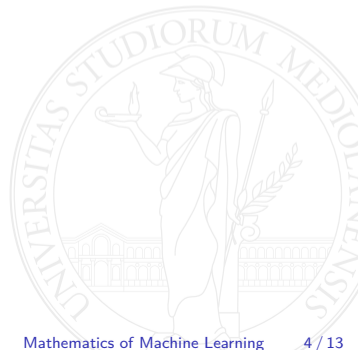
Main idea

- ▶ Control $R_T(\mathbf{u})$ by learning length $w = \|\mathbf{u}\|_2$ and direction $\mathbf{v} = \mathbf{u} / \|\mathbf{u}\|_2$ separately



Main idea

- ▶ Control $R_T(\mathbf{u})$ by learning length $w = \|\mathbf{u}\|_2$ and direction $\mathbf{v} = \mathbf{u} / \|\mathbf{u}\|_2$ separately
- ▶ The **direction** can be learned using FTRL with projection onto the unit Euclidean ball



Main idea

- ▶ Control $R_T(\mathbf{u})$ by learning length $w = \|\mathbf{u}\|_2$ and direction $\mathbf{v} = \mathbf{u} / \|\mathbf{u}\|_2$ separately
- ▶ The **direction** can be learned using FTRL with projection onto the unit Euclidean ball
- ▶ The **length** is learned using a parameterless 1-dimensional online learning algorithm



Main idea

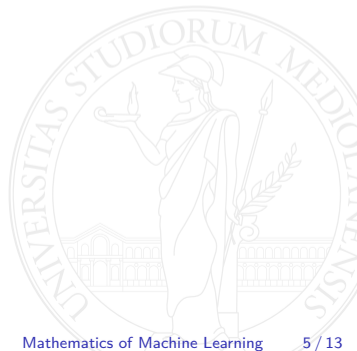
- ▶ Control $R_T(\mathbf{u})$ by learning length $w = \|\mathbf{u}\|_2$ and direction $\mathbf{v} = \mathbf{u} / \|\mathbf{u}\|_2$ separately
- ▶ The **direction** can be learned using FTRL with projection onto the unit Euclidean ball
- ▶ The **length** is learned using a parameterless 1-dimensional online learning algorithm
- ▶ We predict with $w\mathbf{v}$

Analysis



Analysis

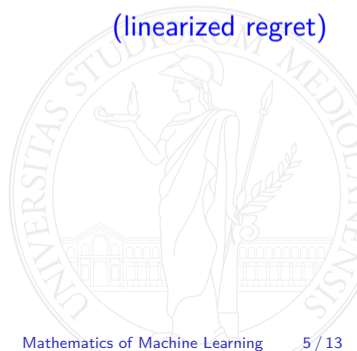
$$R_T(\mathbf{u}) = \sum_{t=1}^T \ell_t(w_t \mathbf{v}_t) - \sum_{t=1}^T \ell_t(\mathbf{u})$$



Analysis

$$\begin{aligned} R_T(\mathbf{u}) &= \sum_{t=1}^T \ell_t(w_t \mathbf{v}_t) - \sum_{t=1}^T \ell_t(\mathbf{u}) \\ &\leq \sum_{t=1}^T \mathbf{g}_t^\top (w_t \mathbf{v}_t - \mathbf{u}) \end{aligned}$$

(linearized regret)



Analysis

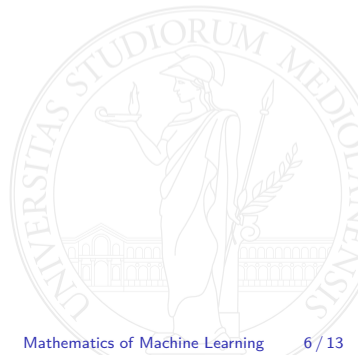
$$\begin{aligned} R_T(\mathbf{u}) &= \sum_{t=1}^T \ell_t(w_t \mathbf{v}_t) - \sum_{t=1}^T \ell_t(\mathbf{u}) \\ &\leq \sum_{t=1}^T \mathbf{g}_t^\top (w_t \mathbf{v}_t - \mathbf{u}) && \text{(linearized regret)} \\ &= \sum_{t=1}^T \left(w_t \mathbf{g}_t^\top \mathbf{v}_t - \|\mathbf{u}\|_2 \mathbf{g}_t^\top \mathbf{v}_t \right) + \|\mathbf{u}\|_2 \sum_{t=1}^T \left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|_2} \right) \end{aligned}$$

Analysis

$$\begin{aligned} R_T(\mathbf{u}) &= \sum_{t=1}^T \ell_t(w_t \mathbf{v}_t) - \sum_{t=1}^T \ell_t(\mathbf{u}) \\ &\leq \sum_{t=1}^T \mathbf{g}_t^\top (w_t \mathbf{v}_t - \mathbf{u}) \quad \text{(linearized regret)} \\ &= \sum_{t=1}^T \underbrace{(w_t \ell'_t(w_t) - \|\mathbf{u}\|_2 \ell'_t(w_t))}_{\text{parameterless}} + \|\mathbf{u}\|_2 \sum_{t=1}^T \underbrace{\left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|_2} \right)}_{\text{FTRL}} \end{aligned}$$

Learning and investing

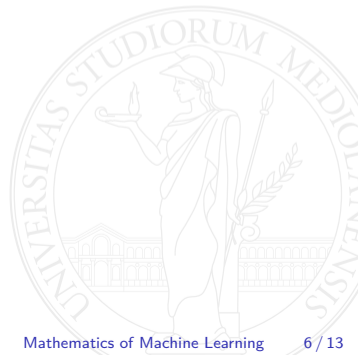
1-dimensional parameterless online algorithms extracted from investment strategies



Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

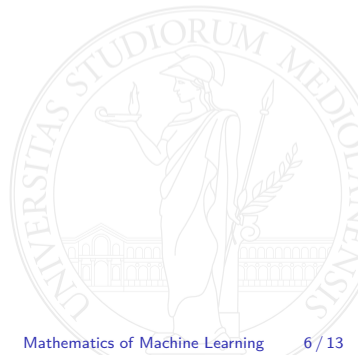


Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$

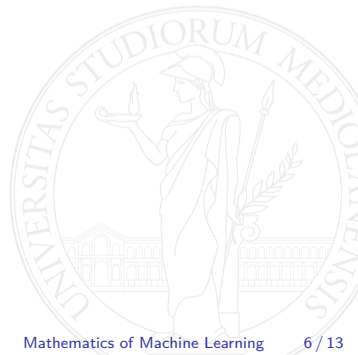


Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game

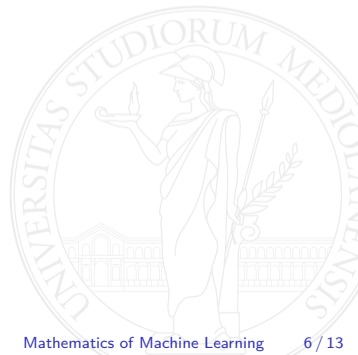


Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$

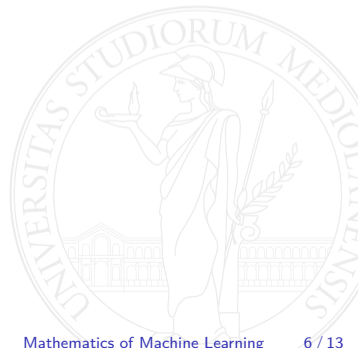


Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$
 2. The market reveals $x_t \in [-1, 1]$

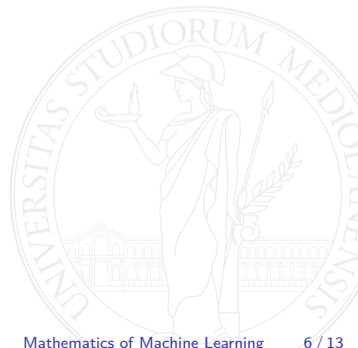


Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$
 2. The market reveals $x_t \in [-1, 1]$
 3. The bettor's wealth is $C_t = (1 + \alpha_t x_t)C_{t-1}$



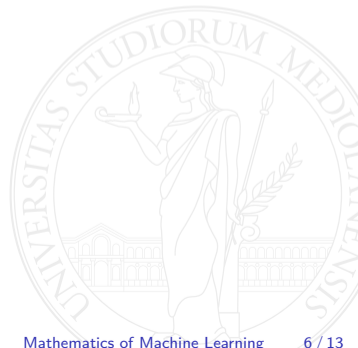
Learning and investing

1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$
 2. The market reveals $x_t \in [-1, 1]$
 3. The bettor's wealth is $C_t = (1 + \alpha_t x_t)C_{t-1}$

A reduction from prediction to investment



Learning and investing

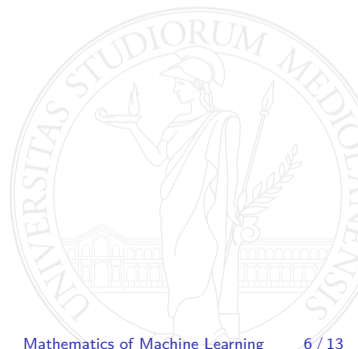
1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$
 2. The market reveals $x_t \in [-1, 1]$
 3. The bettor's wealth is $C_t = (1 + \alpha_t x_t) C_{t-1}$

A reduction from prediction to investment

- ▶ Predict using $w_t = \alpha_t C_{t-1}$ implying $C_t = C_{t-1} + w_t x_t$



Learning and investing

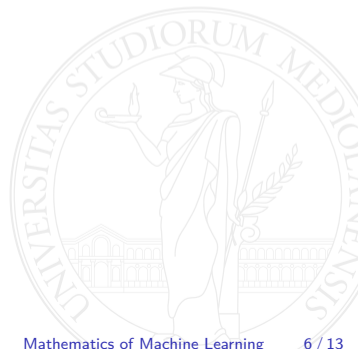
1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$
 2. The market reveals $x_t \in [-1, 1]$
 3. The bettor's wealth is $C_t = (1 + \alpha_t x_t) C_{t-1}$

A reduction from prediction to investment

- ▶ Predict using $w_t = \alpha_t C_{t-1}$ implying $C_t = C_{t-1} + w_t x_t$
- ▶ Provide feedback $x_t = -\ell'_t(w_t) = -\mathbf{g}_t^\top \mathbf{v}_t$



Learning and investing

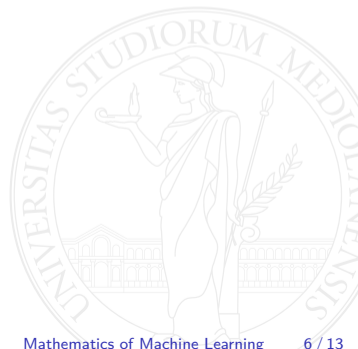
1-dimensional parameterless online algorithms extracted from investment strategies

The betting game

- ▶ The bettor starts out with an initial wealth of $C_0 = 1$
- ▶ In each round $t = 1, 2, \dots$ of the game
 1. The bettor bets $\alpha_t \in [-1, 1]$
 2. The market reveals $x_t \in [-1, 1]$
 3. The bettor's wealth is $C_t = (1 + \alpha_t x_t) C_{t-1}$

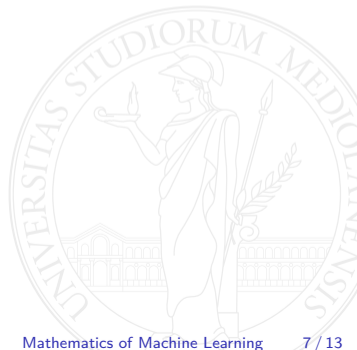
A reduction from prediction to investment

- ▶ Predict using $w_t = \alpha_t C_{t-1}$ implying $C_t = C_{t-1} + w_t x_t$
- ▶ Provide feedback $x_t = -\ell'_t(w_t) = -\mathbf{g}_t^\top \mathbf{v}_t$
- ▶ $C_T = \prod_{t=1}^T (1 + \alpha_t x_t) = 1 + \sum_{t=1}^T w_t x_t = 1 - \sum_{t=1}^T w_t \ell'_t(w_t)$



Connecting wealth and regret

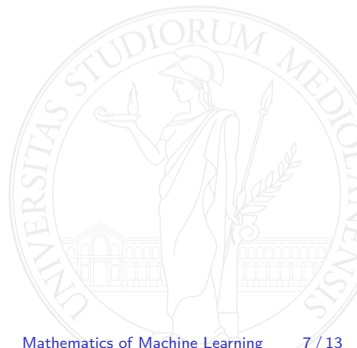
For a convex ϕ assume a betting strategy achieves $C_T \geq \phi \left(\sum_{t=1}^T x_t \right) = \phi \left(- \sum_{t=1}^T \ell'_t(w_t) \right)$



Connecting wealth and regret

For a convex ϕ assume a betting strategy achieves $C_T \geq \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

$$R_T(u) \leq \sum_{t=1}^T (w_t - u) \ell'_t(w_t) \quad (\text{any } u \in \mathbb{R})$$

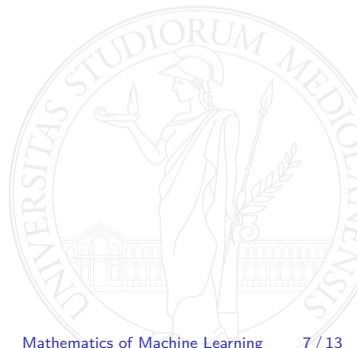


Connecting wealth and regret

For a convex ϕ assume a betting strategy achieves $C_T \geq \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

$$R_T(u) \leq \sum_{t=1}^T (w_t - u) \ell'_t(w_t) \quad (\text{any } u \in \mathbb{R})$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - \left(1 - \sum_{t=1}^T w_t \ell'_t(w_t)\right) + 1$$



Connecting wealth and regret

For a convex ϕ assume a betting strategy achieves $C_T \geq \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

$$R_T(u) \leq \sum_{t=1}^T (w_t - u) \ell'_t(w_t) \quad (\text{any } u \in \mathbb{R})$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - \left(1 - \sum_{t=1}^T w_t \ell'_t(w_t)\right) + 1$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - C_T + 1 \quad (\text{using } C_T = 1 - \sum_{t=1}^T w_t \ell'_t(w_t))$$

Connecting wealth and regret

For a convex ϕ assume a betting strategy achieves $C_T \geq \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

$$R_T(u) \leq \sum_{t=1}^T (w_t - u) \ell'_t(w_t) \quad (\text{any } u \in \mathbb{R})$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - \left(1 - \sum_{t=1}^T w_t \ell'_t(w_t)\right) + 1$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - C_T + 1 \quad (\text{using } C_T = 1 - \sum_{t=1}^T w_t \ell'_t(w_t))$$

$$\leq -u \sum_{t=1}^T \ell'_t(w_t) - \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right) + 1$$

Connecting wealth and regret

For a convex ϕ assume a betting strategy achieves $C_T \geq \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

$$R_T(u) \leq \sum_{t=1}^T (w_t - u) \ell'_t(w_t) \quad (\text{any } u \in \mathbb{R})$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - \left(1 - \sum_{t=1}^T w_t \ell'_t(w_t)\right) + 1$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - C_T + 1 \quad (\text{using } C_T = 1 - \sum_{t=1}^T w_t \ell'_t(w_t))$$

$$\leq -u \sum_{t=1}^T \ell'_t(w_t) - \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right) + 1$$

$$\leq \sup_{\theta \in \mathbb{R}} u\theta - \phi(\theta) + 1 \quad (\theta = -\ell'_1(w_1) - \dots - \ell'_T(w_T))$$

Connecting wealth and regret

For a convex ϕ assume a betting strategy achieves $C_T \geq \phi\left(\sum_{t=1}^T x_t\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

$$R_T(u) \leq \sum_{t=1}^T (w_t - u) \ell'_t(w_t) \quad (\text{any } u \in \mathbb{R})$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - \left(1 - \sum_{t=1}^T w_t \ell'_t(w_t)\right) + 1$$

$$= -u \sum_{t=1}^T \ell'_t(w_t) - C_T + 1 \quad (\text{using } C_T = 1 - \sum_{t=1}^T w_t \ell'_t(w_t))$$

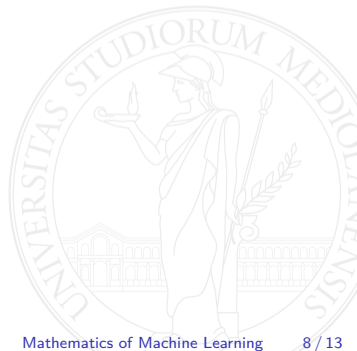
$$\leq -u \sum_{t=1}^T \ell'_t(w_t) - \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right) + 1$$

$$\leq \sup_{\theta \in \mathbb{R}} u\theta - \phi(\theta) + 1 \quad (\theta = -\ell'_1(w_1) - \dots - \ell'_T(w_T))$$

$$= \phi^*(\theta) + 1$$

Regret bound

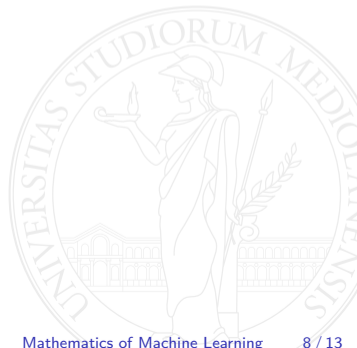
- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)



Regret bound

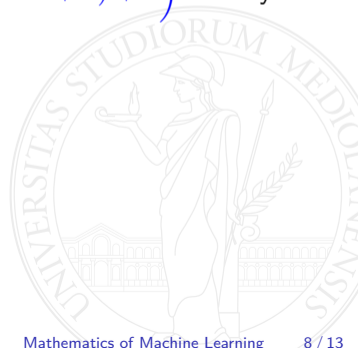
- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)

- ▶ Achieved wealth:
$$C_T \geq \frac{1}{\sqrt{T}} \exp \left(\frac{1}{2T} \left(\sum_{t=1}^T x_t \right)^2 \right) = \phi \left(- \sum_{t=1}^T \ell'_t(w_t) \right)$$



Regret bound

- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)
- ▶ Achieved wealth: $C_T \geq \frac{1}{\sqrt{T}} \exp\left(\frac{1}{2T} \left(\sum_{t=1}^T x_t\right)^2\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$
- ▶ Resulting regret: $R_T(u) = \phi^*\left(-\sum_{t=1}^T \ell'_t(w_t)\right) = \mathcal{O}\left(|u|\sqrt{T \ln(u^2 T + 1)} + 1\right)$ for any $u \in \mathbb{R}$



Regret bound

- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)

- ▶ Achieved wealth: $C_T \geq \frac{1}{\sqrt{T}} \exp\left(\frac{1}{2T} \left(\sum_{t=1}^T x_t\right)^2\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

- ▶ Resulting regret: $R_T(u) = \phi^*\left(-\sum_{t=1}^T \ell'_t(w_t)\right) = \mathcal{O}\left(|u| \sqrt{T \ln(u^2 T + 1)} + 1\right)$ for any

$$u \in \mathbb{R}$$
$$R_T(\mathbf{u}) \leq \sum_{t=1}^T \left(w_t \ell'_t(w_t) - \|\mathbf{u}\|_2 \ell'_t(w_t)\right) + \|\mathbf{u}\|_2 \sum_{t=1}^T \left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|_2}\right) \quad (\text{for any } \mathbf{u} \in \mathbb{R}^d)$$

Regret bound

- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)

- ▶ Achieved wealth: $C_T \geq \frac{1}{\sqrt{T}} \exp\left(\frac{1}{2T} \left(\sum_{t=1}^T x_t\right)^2\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

- ▶ Resulting regret: $R_T(u) = \phi^*\left(-\sum_{t=1}^T \ell'_t(w_t)\right) = \mathcal{O}\left(|u| \sqrt{T \ln(u^2 T + 1)} + 1\right)$ for any

$$\begin{aligned} u \in \mathbb{R} \\ R_T(\mathbf{u}) &\leq \sum_{t=1}^T \left(w_t \ell'_t(w_t) - \|\mathbf{u}\|_2 \ell'_t(w_t)\right) + \|\mathbf{u}\|_2 \sum_{t=1}^T \left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|_2}\right) \quad (\text{for any } \mathbf{u} \in \mathbb{R}^d) \\ &= \mathcal{O}\left(\|\mathbf{u}\|_2 \sqrt{T \ln(\|\mathbf{u}\|_2^2 T + 1)} + 1\right) + \mathcal{O}(\|\mathbf{u}\|_2 \sqrt{T}) \end{aligned}$$

Regret bound

- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)

- ▶ Achieved wealth: $C_T \geq \frac{1}{\sqrt{T}} \exp\left(\frac{1}{2T} \left(\sum_{t=1}^T x_t\right)^2\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

- ▶ Resulting regret: $R_T(u) = \phi^*\left(-\sum_{t=1}^T \ell'_t(w_t)\right) = \mathcal{O}\left(|u| \sqrt{T \ln(u^2 T + 1)} + 1\right)$ for any

$$\begin{aligned} u &\in \mathbb{R} \\ R_T(\mathbf{u}) &\leq \sum_{t=1}^T \left(w_t \ell'_t(w_t) - \|\mathbf{u}\|_2 \ell'_t(w_t)\right) + \|\mathbf{u}\|_2 \sum_{t=1}^T \left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|_2}\right) \quad (\text{for any } \mathbf{u} \in \mathbb{R}^d) \\ &= \mathcal{O}\left(\|\mathbf{u}\|_2 \sqrt{T \ln(\|\mathbf{u}\|_2^2 T + 1)} + 1\right) + \mathcal{O}(\|\mathbf{u}\|_2 \sqrt{T}) \\ &= \mathcal{O}\left(\|\mathbf{u}\|_2 \sqrt{T \ln(\|\mathbf{u}\|_2^2 T + 1)} + 1\right) \end{aligned}$$

Regret bound

- ▶ Betting strategy: $\alpha_1 = 0$ and $\alpha_t = (x_1 + \dots + x_{t-1})/t$ for $t \geq 1$ (Krichevsky-Trofimov estimator)

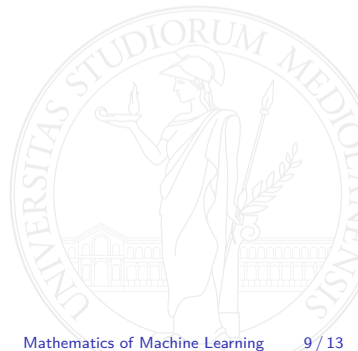
- ▶ Achieved wealth: $C_T \geq \frac{1}{\sqrt{T}} \exp\left(\frac{1}{2T} \left(\sum_{t=1}^T x_t\right)^2\right) = \phi\left(-\sum_{t=1}^T \ell'_t(w_t)\right)$

- ▶ Resulting regret: $R_T(u) = \phi^*\left(-\sum_{t=1}^T \ell'_t(w_t)\right) = \mathcal{O}\left(|u| \sqrt{T \ln(u^2 T + 1)} + 1\right)$ for any

$$\begin{aligned} & u \in \mathbb{R} \\ R_T(\mathbf{u}) & \leq \sum_{t=1}^T \left(w_t \ell'_t(w_t) - \|\mathbf{u}\|_2 \ell'_t(w_t) \right) + \|\mathbf{u}\|_2 \sum_{t=1}^T \left(\mathbf{g}_t^\top \mathbf{v}_t - \mathbf{g}_t^\top \frac{\mathbf{u}}{\|\mathbf{u}\|_2} \right) \quad (\text{for any } \mathbf{u} \in \mathbb{R}^d) \\ & = \mathcal{O}\left(\|\mathbf{u}\|_2 \sqrt{T \ln(\|\mathbf{u}\|_2^2 T + 1)} + 1\right) + \mathcal{O}(\|\mathbf{u}\|_2 \sqrt{T}) \\ & = \mathcal{O}\left(\|\mathbf{u}\|_2 \sqrt{T \ln(\|\mathbf{u}\|_2^2 T + 1)} + 1\right) \end{aligned}$$

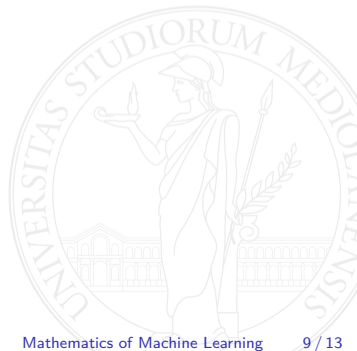
Result matches the $\|\mathbf{u}\|_2 \sqrt{T}$ bound up to log factors that are unavoidable if $\|\mathbf{u}\|_2$ is unknown

Other notions of regret



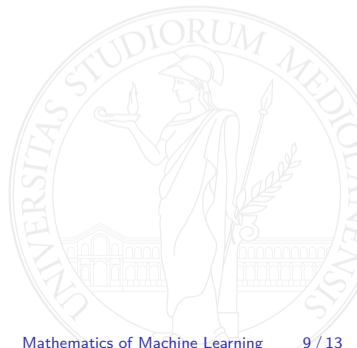
Other notions of regret

- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless



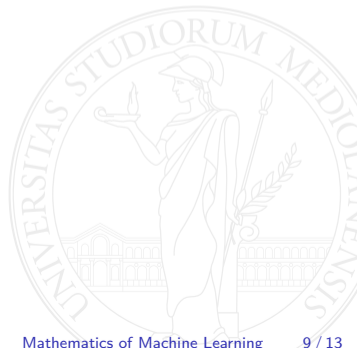
Other notions of regret

- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence



Other notions of regret

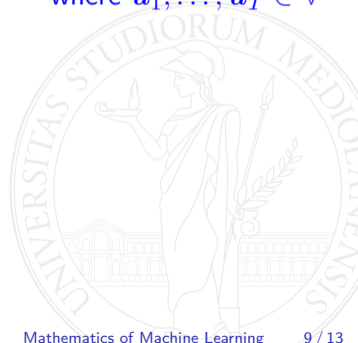
- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence
- ▶ In this case, the regret should be replaced by more robust measures



Other notions of regret

- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence
- ▶ In this case, the regret should be replaced by more robust measures

- ▶ **Dynamic regret** $R_T^{\text{dyn}}(\mathbf{u}_1, \dots, \mathbf{u}_T) = \sum_{t=1}^T l_t(\mathbf{w}_t) - \sum_{t=1}^T l_t(\mathbf{u}_t)$ where $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathbb{V}$

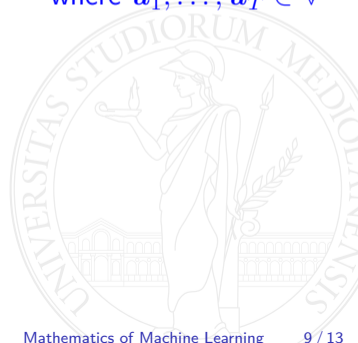


Other notions of regret

- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence
- ▶ In this case, the regret should be replaced by more robust measures

▶ **Dynamic regret** $R_T^{\text{dyn}}(\mathbf{u}_1, \dots, \mathbf{u}_T) = \sum_{t=1}^T l_t(\mathbf{w}_t) - \sum_{t=1}^T l_t(\mathbf{u}_t)$ where $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathbb{V}$

▶ Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|$



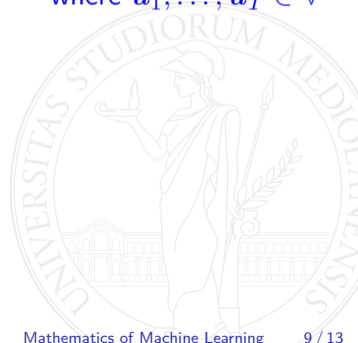
Other notions of regret

- ▶ If the loss sequence ℓ_1, ℓ_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $\ell_1(\mathbf{u}) + \ell_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence
- ▶ In this case, the regret should be replaced by more robust measures

▶ **Dynamic regret** $R_T^{\text{dyn}}(\mathbf{u}_1, \dots, \mathbf{u}_T) = \sum_{t=1}^T \ell_t(\mathbf{w}_t) - \sum_{t=1}^T \ell_t(\mathbf{u}_t)$ where $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathbb{V}$

▶ Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|$

▶ Lower bound: $\Omega(G\sqrt{(D + \Pi_T)DT})$



Other notions of regret

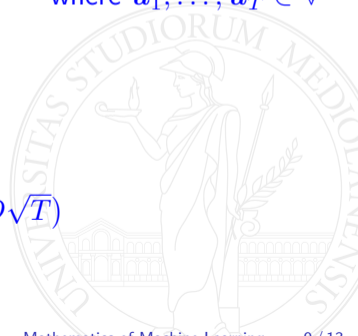
- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence
- ▶ In this case, the regret should be replaced by more robust measures

▶ **Dynamic regret** $R_T^{\text{dyn}}(\mathbf{u}_1, \dots, \mathbf{u}_T) = \sum_{t=1}^T l_t(\mathbf{w}_t) - \sum_{t=1}^T l_t(\mathbf{u}_t)$ where $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathbb{V}$

▶ Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|$

▶ Lower bound: $\Omega(G\sqrt{(D + \Pi_T)DT})$

▶ When $\Pi_T = 0$ this reduces to the standard lower bound $\Omega(GD\sqrt{T})$



Other notions of regret

- ▶ If the loss sequence l_1, l_2, \dots is such that no $\mathbf{u} \in \mathbb{V}$ achieves a small cumulative loss $l_1(\mathbf{u}) + l_2(\mathbf{u}) + \dots$, then regret bounds are meaningless
- ▶ Lack of a single good minimizer in \mathbb{V} caused by a highly nonstationary data sequence
- ▶ In this case, the regret should be replaced by more robust measures

▶ **Dynamic regret** $R_T^{\text{dyn}}(\mathbf{u}_1, \dots, \mathbf{u}_T) = \sum_{t=1}^T l_t(\mathbf{w}_t) - \sum_{t=1}^T l_t(\mathbf{u}_t)$ where $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathbb{V}$

▶ Complexity parameter: $\Pi_T = \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|$

▶ Lower bound: $\Omega(G\sqrt{(D + \Pi_T)DT})$

▶ When $\Pi_T = 0$ this reduces to the standard lower bound $\Omega(GD\sqrt{T})$

- ▶ Matching upper bound obtained by using Hedge to aggregate $\mathcal{O}(\ln T)$ instances of FTRL each tuned to a different value of Π_T

Adaptive regret



Adaptive regret

- ▶ Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time



Adaptive regret

- ▶ Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time

- ▶ $R_{\tau, T}^{\text{ada}} = \max_{s=1, \dots, T-\tau+1} \left(\sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{w}_t) - \min_{\mathbf{u} \in \mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{u}) \right)$

where $\tau \in \{1, \dots, T\}$



Adaptive regret

- ▶ Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time

- ▶ $R_{\tau,T}^{\text{ada}} = \max_{s=1,\dots,T-\tau+1} \left(\sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{w}_t) - \min_{\mathbf{u} \in \mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{u}) \right)$ where $\tau \in \{1, \dots, T\}$

- ▶ Best known upper bound: $R_{\tau,T}^{\text{ada}}(\mathbf{u}) = \mathcal{O}(DG\sqrt{\tau} + \sqrt{(\ln T)\tau})$

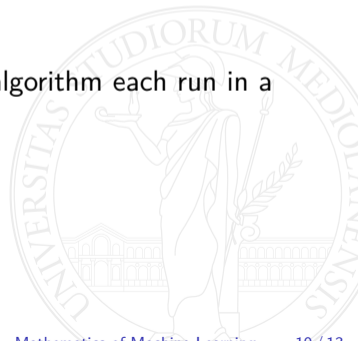


Adaptive regret

- ▶ Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time

- ▶ $R_{\tau,T}^{\text{ada}} = \max_{s=1,\dots,T-\tau+1} \left(\sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{w}_t) - \min_{\mathbf{u} \in \mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{u}) \right)$ where $\tau \in \{1, \dots, T\}$

- ▶ Best known upper bound: $R_{\tau,T}^{\text{ada}}(\mathbf{u}) = \mathcal{O}(DG\sqrt{\tau} + \sqrt{(\ln T)\tau})$
- ▶ Obtained by combining several instances of a standard online algorithm each run in a specific interval of time



Adaptive regret

- ▶ Evaluate the performance of the online algorithm against that of the best fixed comparator in any interval of time

- ▶ $R_{\tau,T}^{\text{ada}} = \max_{s=1,\dots,T-\tau+1} \left(\sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{w}_t) - \min_{\mathbf{u} \in \mathbb{V}} \sum_{t=s}^{s+\tau-1} \ell_t(\mathbf{u}) \right)$ where $\tau \in \{1, \dots, T\}$

- ▶ Best known upper bound: $R_{\tau,T}^{\text{ada}}(\mathbf{u}) = \mathcal{O}(DG\sqrt{\tau} + \sqrt{(\ln T)\tau})$
- ▶ Obtained by combining several instances of a standard online algorithm each run in a specific interval of time
- ▶ The set of intervals is carefully designed so that the overall number of instances to be run is $\mathcal{O}(\ln T)$

From sequential to statistical learning

- ▶ Statistical risk for a convex and bounded loss $\ell_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}^{\top} \mathbf{X}, Y)]$



From sequential to statistical learning

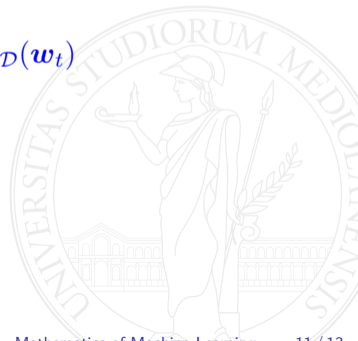
- ▶ Statistical risk for a convex and bounded loss $\ell_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}^\top \mathbf{X}, Y)]$
- ▶ Let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ where $\mathbf{w}_1, \dots, \mathbf{w}_T$ are generated by an online algorithm over $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ drawn i.i.d. from an unknown distribution \mathcal{D}



From sequential to statistical learning

- ▶ Statistical risk for a convex and bounded loss $\ell_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}^{\top} \mathbf{X}, Y)]$
- ▶ Let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ where $\mathbf{w}_1, \dots, \mathbf{w}_T$ are generated by an online algorithm over $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ drawn i.i.d. from an unknown distribution \mathcal{D}
- ▶ By Jensen's inequality

$$\ell_{\mathcal{D}}(\bar{\mathbf{w}}) = \mathbb{E}[\ell(\bar{\mathbf{w}}^{\top} \mathbf{X}, Y)] \leq \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \ell(\mathbf{w}_t^{\top} \mathbf{X}_t, Y_t)\right] = \frac{1}{T} \sum_{t=1}^T \ell_{\mathcal{D}}(\mathbf{w}_t)$$



From sequential to statistical learning

- ▶ Statistical risk for a convex and bounded loss $l_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}^{\top} \mathbf{X}, Y)]$
- ▶ Let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ where $\mathbf{w}_1, \dots, \mathbf{w}_T$ are generated by an online algorithm over $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ drawn i.i.d. from an unknown distribution \mathcal{D}

- ▶ By Jensen's inequality

$$l_{\mathcal{D}}(\bar{\mathbf{w}}) = \mathbb{E}[\ell(\bar{\mathbf{w}}^{\top} \mathbf{X}, Y)] \leq \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \ell(\mathbf{w}_t^{\top} \mathbf{X}_t, Y_t)\right] = \frac{1}{T} \sum_{t=1}^T l_{\mathcal{D}}(\mathbf{w}_t)$$

- ▶ Note also that $\mathbb{E}[l_{\mathcal{D}}(\mathbf{w}_t) - \ell(\mathbf{w}_t^{\top} \mathbf{X}_t, Y_t) \mid (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{t-1}, Y_{t-1})] = 0$

From sequential to statistical learning

- ▶ Statistical risk for a convex and bounded loss $\ell_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}^{\top} \mathbf{X}, Y)]$
- ▶ Let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ where $\mathbf{w}_1, \dots, \mathbf{w}_T$ are generated by an online algorithm over $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ drawn i.i.d. from an unknown distribution \mathcal{D}

- ▶ By Jensen's inequality

$$\ell_{\mathcal{D}}(\bar{\mathbf{w}}) = \mathbb{E}[\ell(\bar{\mathbf{w}}^{\top} \mathbf{X}, Y)] \leq \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \ell(\mathbf{w}_t^{\top} \mathbf{X}_t, Y_t)\right] = \frac{1}{T} \sum_{t=1}^T \ell_{\mathcal{D}}(\mathbf{w}_t)$$

- ▶ Note also that $\mathbb{E}[\ell_{\mathcal{D}}(\mathbf{w}_t) - \ell(\mathbf{w}_t^{\top} \mathbf{X}_t, Y_t) \mid (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{t-1}, Y_{t-1})] = 0$
- ▶ Using concentration inequalities for martingales (e.g., Hoeffding-Azuma inequality),

$$\frac{1}{T} \sum_{t=1}^T \ell_{\mathcal{D}}(\mathbf{w}_t) \leq \frac{1}{T} \sum_{t=1}^T \ell(\mathbf{w}_t^{\top} \mathbf{X}_t, Y_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

w.h.p.

Statistical risk bounds from regret bounds

Letting $l(\mathbf{w}^\top \mathbf{X}_t, Y_t) = \ell_t(\mathbf{w})$ we have $l_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ w.h.p.



Statistical risk bounds from regret bounds

Letting $l(\mathbf{w}^\top \mathbf{X}_t, Y_t) = \ell_t(\mathbf{w})$ we have $l_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ w.h.p.

Let $\mathbf{u} = \underset{\mathbf{w} \in \mathcal{V}}{\operatorname{argmin}} \ell_{\mathcal{D}}(\mathbf{w})$ and bound the estimation error $l_{\mathcal{D}}(\bar{\mathbf{w}}) - \ell_{\mathcal{D}}(\mathbf{u})$ w.h.p.

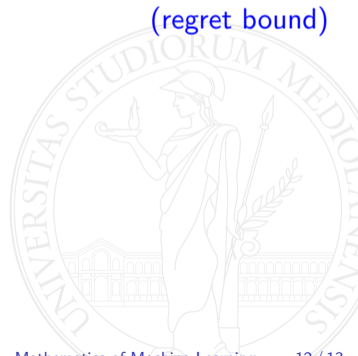


Statistical risk bounds from regret bounds

Letting $l(\mathbf{w}^\top \mathbf{X}_t, Y_t) = \ell_t(\mathbf{w})$ we have $l_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ w.h.p.

Let $\mathbf{u} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} l_{\mathcal{D}}(\mathbf{w})$ and bound the estimation error $l_{\mathcal{D}}(\bar{\mathbf{w}}) - l_{\mathcal{D}}(\mathbf{u})$ w.h.p.

$$\frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) \leq \frac{1}{T} \inf_{\mathbf{w} \in \mathcal{V}} \sum_{t=1}^T \ell_t(\mathbf{w}) + \frac{2GD}{\sqrt{T}} \quad (\text{regret bound})$$



Statistical risk bounds from regret bounds

Letting $l(\mathbf{w}^\top \mathbf{X}_t, Y_t) = \ell_t(\mathbf{w})$ we have $l_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ w.h.p.

Let $\mathbf{u} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} l_{\mathcal{D}}(\mathbf{w})$ and bound the estimation error $l_{\mathcal{D}}(\bar{\mathbf{w}}) - l_{\mathcal{D}}(\mathbf{u})$ w.h.p.

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) &\leq \frac{1}{T} \inf_{\mathbf{w} \in \mathcal{V}} \sum_{t=1}^T \ell_t(\mathbf{w}) + \frac{2GD}{\sqrt{T}} && \text{(regret bound)} \\ &\leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u}) + \frac{2GD}{\sqrt{T}} \end{aligned}$$

Statistical risk bounds from regret bounds

Letting $l(\mathbf{w}^\top \mathbf{X}_t, Y_t) = \ell_t(\mathbf{w})$ we have $l_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ w.h.p.

Let $\mathbf{u} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{V}} l_{\mathcal{D}}(\mathbf{w})$ and bound the estimation error $l_{\mathcal{D}}(\bar{\mathbf{w}}) - l_{\mathcal{D}}(\mathbf{u})$ w.h.p.

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) &\leq \frac{1}{T} \inf_{\mathbf{w} \in \mathcal{V}} \sum_{t=1}^T \ell_t(\mathbf{w}) + \frac{2GD}{\sqrt{T}} && \text{(regret bound)} \\ &\leq \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u}) + \frac{2GD}{\sqrt{T}} \\ &\leq l_{\mathcal{D}}(\mathbf{u}) + \frac{2GD}{\sqrt{T}} + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) && \text{w.h.p.} \end{aligned}$$

using concentration of $\ell_t(\mathbf{u})$ around $l_{\mathcal{D}}(\mathbf{u})$

Final bound

If $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ where $\mathbf{w}_1, \dots, \mathbf{w}_T$ are generated by an online algorithm over $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ drawn i.i.d. from an unknown distribution \mathcal{D} , then

$$\ell_{\mathcal{D}}(\bar{\mathbf{w}}) - \inf_{\mathbf{w} \in \mathcal{V}} \ell_{\mathcal{D}}(\mathbf{w}) \leq \frac{2GD}{\sqrt{T}} \quad \text{w.h.p.}$$

