The Many Faces of Exponential Weights in Online Learning

Tim van Erven

Universiteit Leiden

Joint work with:
Dirk van der Hoeven, Wojciech Kotłowski, Wouter Koolen

CWI-Inria Workshop, September 17, 2019

Example: Betting on Football Games

Netherlands runner-up in Women's World Cup 2019

- Before every match t in the English Premier League, my co-author Dirk wants to predict the goal difference Y_t
- Given feature vector $\boldsymbol{X}_t \in \mathbb{R}^d$, he may predict $\hat{Y}_t = \boldsymbol{X}_t^\intercal \boldsymbol{\theta}_t$ with a linear model
- After the match: observe Y_t
- Measure **loss** by $f_t(\boldsymbol{\theta}_t) = (Y_t - \hat{Y}_t)^2$ and improve **parameter estimates**: $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1}$

Example: Betting on Football Games

Netherlands runner-up in Women's World Cup 2019

- Before every match t in the English Premier League, my co-author Dirk wants to predict the goal difference Y_t
- Given feature vector $\boldsymbol{X}_t \in \mathbb{R}^d$, he may predict $\hat{Y}_t = \boldsymbol{X}_t^\intercal \boldsymbol{\theta}_t$ with a linear model
- After the match: observe Y_t
- Measure **loss** by $f_t(\boldsymbol{\theta}_t) = (Y_t - \hat{Y}_t)^2$ and improve **parameter estimates**: $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1}$

Goal: Predict almost as well as the best possible parameters $\boldsymbol{\theta}^*$:

$$\mathbf{Regret}_T(\boldsymbol{\theta}^*) = \sum_{t=1}^{T} f_t(\boldsymbol{\theta}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{\theta}^*)$$

Online Convex Optimization

Parameters $\boldsymbol{\theta}$ take values in a convex domain $\Theta \subset \mathbb{R}^d$

for $t = 1, 2, \ldots, T$ **do**
    Learner **estimates** $\boldsymbol{\theta}_t \in \Theta$
    Nature reveals **convex loss function** $f_t : \Theta \to \mathbb{R}$
end for

Viewed as a **zero-sum game** against Nature:

$$V \;=\; \min_{\boldsymbol{\theta}_1} \max_{f_1} \min_{\boldsymbol{\theta}_2} \max_{f_2} \cdots \min_{\boldsymbol{\theta}_T} \max_{f_T} \max_{\boldsymbol{\theta}^* \in \Theta} \; \mathrm{Regret}_T(\boldsymbol{\theta}^*)$$

Online Gradient Descent

$$\tilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla f_t(\boldsymbol{\theta}_t)$$
$$\boldsymbol{\theta}_{t+1} = \min_{\boldsymbol{\theta}\in\Theta} \|\tilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}\|$$

Theorem (Zinkevich, 2003)

Suppose Θ compact with diameter at most D, and $\|\nabla f_t(\boldsymbol{\theta}_t)\|_2 \le G$. Then online gradient descent with $\eta_t = \frac{D}{G\sqrt{t}}$ guarantees

$$\mathrm{Regret}_T(\boldsymbol{\theta}^*) \le \frac{3}{2} GD\sqrt{T}$$

for **any** *choices of Nature.*

Under these assumptions, this is optimal (up to a constant factor).

Example: Prediction with Expert Advice

	Day 1	**Day 2**	**Day 3**	$\cdots$	**Day T**
Expert 1				$\cdots$	
Expert 2				$\cdots$	
Expert 3				$\cdots$	
Truth				$\cdots$	

Example: Prediction with Expert Advice

	Day 1	**Day 2**	**Day 3**	⋯	**Day T**
Expert 1				⋯	
Expert 2				⋯	
Expert 3				⋯	
Truth				⋯	

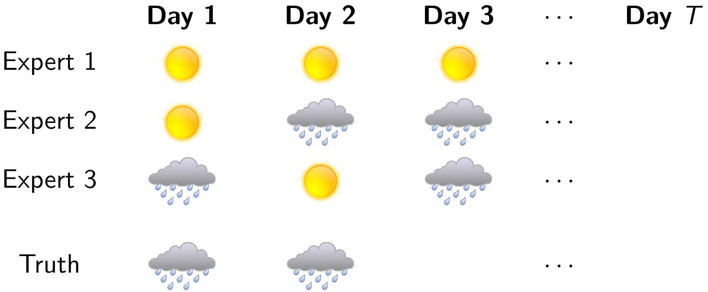
Example: Prediction with Expert Advice

	Day 1	**Day 2**	**Day 3**	$\cdots$	**Day T**
Expert 1				$\cdots$	
Expert 2				$\cdots$	
Expert 3				$\cdots$	
Truth				$\cdots$	

Example: Prediction with Expert Advice

	Day 1	**Day 2**	**Day 3**	$\cdots$	**Day T**
Expert 1				$\cdots$	
Expert 2				$\cdots$	
Expert 3				$\cdots$	
Truth				$\cdots$	

Example: Prediction with Expert Advice



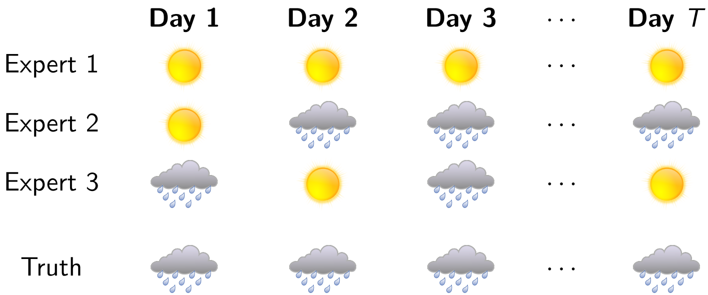
Example: Prediction with Expert Advice

	Day 1	**Day 2**	**Day 3**	$\cdots$	**Day T**
Expert 1				$\cdots$	
Expert 2				$\cdots$	
Expert 3				$\cdots$	
Truth				$\cdots$	

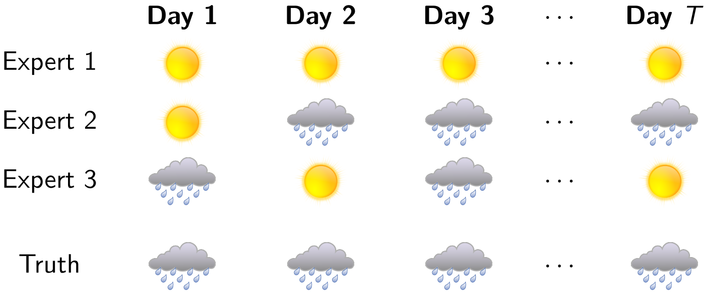
Example: Prediction with Expert Advice

	Day 1	**Day 2**	**Day 3**	$\cdots$	**Day T**
Expert 1				$\cdots$	
Expert 2				$\cdots$	
Expert 3				$\cdots$	
Truth				$\cdots$	

Example: Prediction with Expert Advice



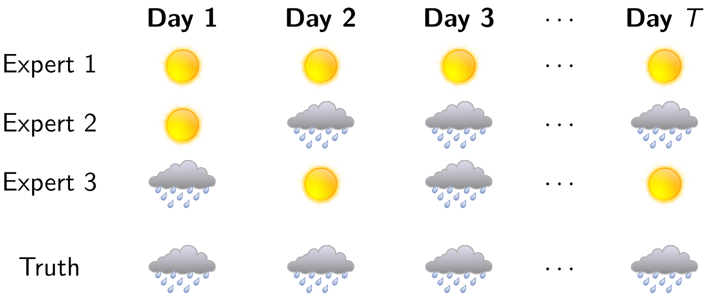
Example: Prediction with Expert Advice



Fits in Framework:

- Linear loss: $f_t(\boldsymbol{\theta}) = \boldsymbol{g}_t^\intercal \boldsymbol{\theta}$
  where $\boldsymbol{g}_t \in \{0, 1\}^d$ contains mistakes of d experts
- Compare with deterministic choice of expert $\boldsymbol{\theta}^* \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$
- But allow randomized predictions: $\boldsymbol{\theta}_t = \mathbb{E}_{P_t(i)}[\boldsymbol{e}_i]$

Example: Prediction with Expert Advice



Fits in Framework:

- Linear loss: $f_t(\boldsymbol{\theta}) = \boldsymbol{g}_t^\intercal \boldsymbol{\theta}$
  where $\boldsymbol{g}_t \in \{0,1\}^d$ contains mistakes of d experts
- Compare with deterministic choice of expert $\boldsymbol{\theta}^* \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$
- But allow randomized predictions: $\boldsymbol{\theta}_t = \mathbb{E}_{P_t(i)}[\boldsymbol{e}_i]$

GD Regret Bound	**Optimal Regret Bound**
$O(GD\sqrt{T}) = O(\sqrt{d\,T})$	$O(\sqrt{\log(d)\,T})$

Exponential Weights for Expert Advice

- Given **prior distribution** P_1 on experts $\{1, \ldots, d\}$
- Choose expert i with probability

$$P_{t+1}(i) = \frac{\exp\left(-\eta_t \sum_{s=1}^{t} g_{s,i}\right) P_1(i)}{\sum_{j=1}^{d} \exp\left(-\eta_t \sum_{s=1}^{t} g_{s,j}\right) P_1(j)}$$

Exponential Weights for Expert Advice

- Given **prior distribution** P_1 on experts $\{1, \ldots, d\}$
- Choose expert i with probability

$$P_{t+1}(i) = \frac{\exp\left(-\eta_t \sum_{s=1}^{t} g_{s,i}\right) P_1(i)}{\sum_{j=1}^{d} \exp\left(-\eta_t \sum_{s=1}^{t} g_{s,j}\right) P_1(j)}$$

Theorem (Vovk, 1990, Littlestone, Warmuth, 1994)

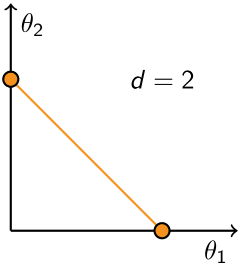
Exponential weights for expert advice with uniform prior P_1 and $\eta_t = \sqrt{\frac{8 \log(d)}{T}}$ guarantees

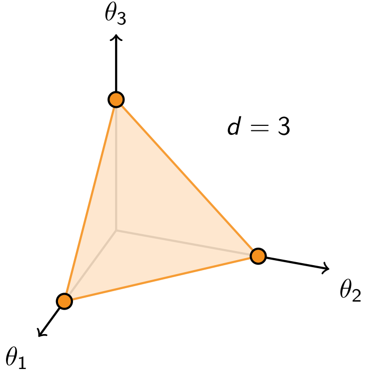
$$\text{Regret}_T(\boldsymbol{\theta}^*) \leq \sqrt{\tfrac{1}{2} \log(d)\, T}$$

for ***any*** *choices of Nature.*

This is optimal for experts (with exactly these constants).

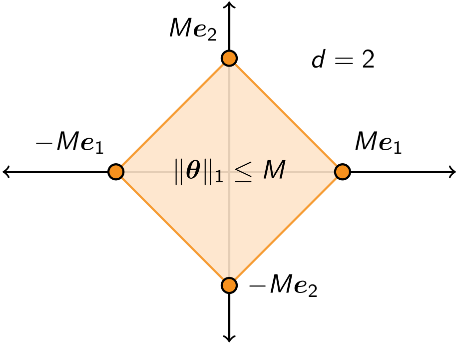
A Broader View of Exponential Weights





- Linear loss: $f_t(\boldsymbol{\theta}) = \boldsymbol{g}_t^\intercal \boldsymbol{\theta}$
- Prior P_1 supported on corners of simplex $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$
- Distribution of predictions is mean of P_t: $\boldsymbol{\theta}_t = \mathbb{E}_{P_t(\boldsymbol{\theta})}[\boldsymbol{\theta}]$

Scaling up Exponential Weights



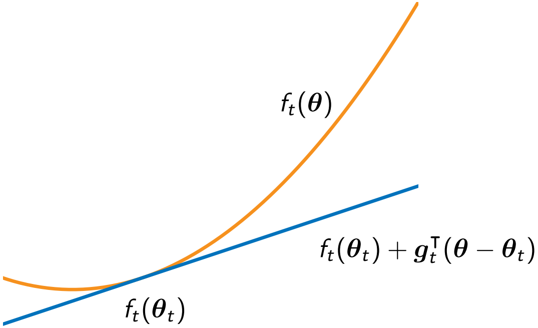
Theorem ($\mathrm{EG}^{\pm}$, Kivinen, Warmuth, 1997)

Suppose $f_t(\theta) = g_t^\intercal \theta$ and $\|g_t\|_\infty \leq G$. Then exponential weights with uniform prior on $\{\pm M e_1, \ldots, \pm M e_d\}$ and $\eta_t = \sqrt{\frac{2\log(2d)}{M^2 G^2 T}}$ guarantees

$$\mathrm{Regret}_T(\theta^*) \leq GM\sqrt{2\log(2d)T}$$

for all θ^* with $\|\theta^*\|_1 \leq M$.

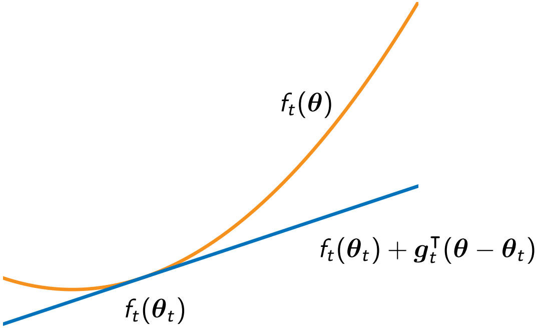
Linearizing General Convex Losses



For $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{\theta}_t)$, the linear loss $\tilde{f}_t(\boldsymbol{\theta}) = \boldsymbol{g}_t^\intercal \boldsymbol{\theta}$ satisfies

$$f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}^*) \leq \boldsymbol{g}_t^\intercal(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) = \tilde{f}_t(\boldsymbol{\theta}_t) - \tilde{f}_t(\boldsymbol{\theta}^*)$$

Linearizing General Convex Losses

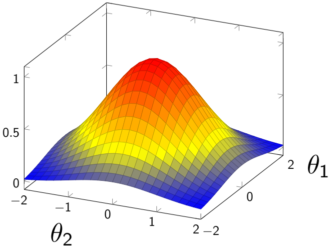


For $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{\theta}_t)$, the linear loss $\tilde{f}_t(\boldsymbol{\theta}) = \boldsymbol{g}_t^\intercal \boldsymbol{\theta}$ satisfies

$$f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}^*) \leq \boldsymbol{g}_t^\intercal(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) = \tilde{f}_t(\boldsymbol{\theta}_t) - \tilde{f}_t(\boldsymbol{\theta}^*)$$

- To prevent infinite regret, need $|\tilde{f}_t(\boldsymbol{\theta})|$ to be bounded.
- Hence dual norms to bound domain and gradients: $|\tilde{f}_t(\boldsymbol{\theta})| \leq \|\boldsymbol{g}_t\|_p \cdot \|\boldsymbol{\theta}\|_q$ for $1/p + 1/q = 1$

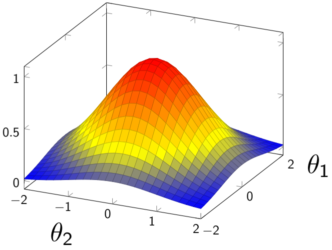
Exponential Weights for L_2-Domains



- **Multivariate Gaussian** prior: $P_1 = \mathcal{N}(0, \boldsymbol{I})$
- **Linearized losses**

$$\mathrm{d}P_{t+1}(\boldsymbol{\theta}) \propto \exp\left(-\eta_t \sum_{s=1}^{t} \boldsymbol{g}_s^\intercal \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^\intercal \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta}$$

Exponential Weights for L_2-Domains

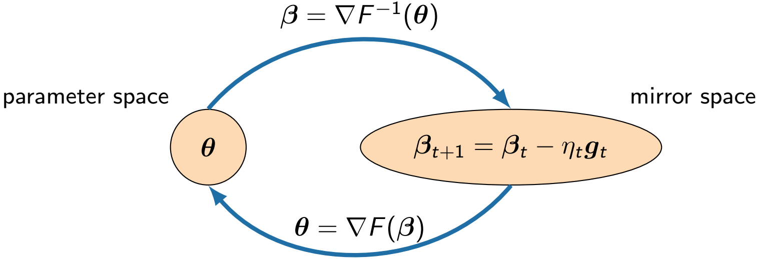


- **Multivariate Gaussian** prior: $P_1 = \mathcal{N}(0, \boldsymbol{I})$
- **Linearized losses**

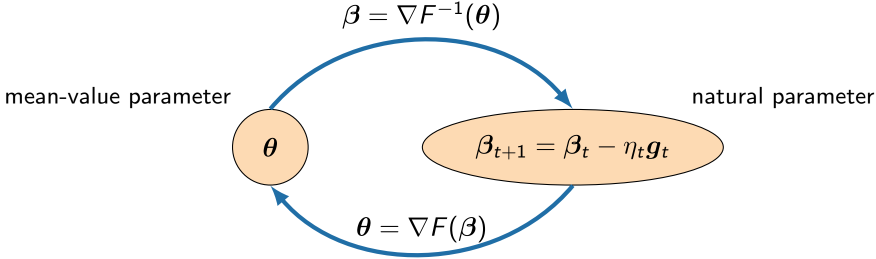
$$\mathrm{d}P_{t+1}(\boldsymbol{\theta}) \propto \exp\left(-\eta_t \sum_{s=1}^{t} \boldsymbol{g}_s^\intercal \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^\intercal \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta}$$

Recover gradient descent! $P_{t+1} = \mathcal{N}(-\eta_t \sum_{s=1}^{t} \boldsymbol{g}_s, \boldsymbol{I})$

Mirror Descent



Mirror Descent



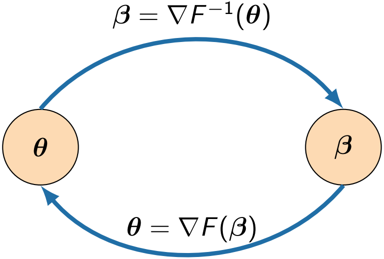
Theorem

Mirror descent is the mean of exponential weights with prior P_1 if $\{\boldsymbol{\beta} : F(\boldsymbol{\beta}) < \infty\}$ is an open set and

$$F(\boldsymbol{\beta}) = \ln \int e^{\boldsymbol{\beta}^\intercal \boldsymbol{\theta}} \mathrm{d}P_1(\boldsymbol{\theta}).$$

Interpretation: F is the cumulant generating function of a regular exponential family with carrier P_1.

Mirror Descent



Examples:

- **Gradient descent**:
  $F(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$, $P_1 = \mathcal{N}(0, \boldsymbol{I})$
- **Unnormalized relative entropy**:
  $F(\boldsymbol{\beta}) = \sum_{i=1}^{d} e^{\beta_i}$,
  $P_1 = \prod_{i=1}^{d} P_{\lambda_i}(\theta_i)$ is product of Poisson distributions

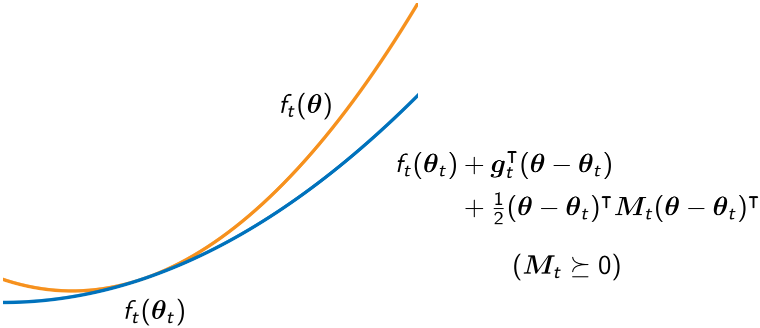
Theorem

Mirror descent is the mean of exponential weights with prior P_1 if $\{\boldsymbol{\beta} : F(\boldsymbol{\beta}) < \infty\}$ is an open set and

$$F(\boldsymbol{\beta}) = \ln \int e^{\boldsymbol{\beta}^\intercal \boldsymbol{\theta}} \mathrm{d}P_1(\boldsymbol{\theta}).$$

Interpretation: F is the cumulant generating function of a regular exponential family with carrier P_1.

Quadratic Lower-bounded Losses



The quadratic loss $\tilde{f}_t(\boldsymbol{\theta}) = \boldsymbol{g}_t^\intercal \boldsymbol{\theta} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\intercal \boldsymbol{M}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\intercal$ satisfies

$$f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}^*) \leq \tilde{f}_t(\boldsymbol{\theta}_t) - \tilde{f}_t(\boldsymbol{\theta}^*)$$

EW for Quadratic Lower-bounded Losses

Theorem

Exponential weights with Gaussian prior $P_1 = \mathcal{N}(0, \boldsymbol{I})$ *and constant* $\eta_t = \eta$ *produces Gaussian distributions* $P_{t+1} = \mathcal{N}(\boldsymbol{\theta}_{t+1}, \Sigma_{t+1})$ *and guarantees*

$$\mathrm{Regret}_T(\boldsymbol{\theta}^*) \leq \frac{1}{2\eta}\|\boldsymbol{\theta}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \Sigma_{t+1} \boldsymbol{g}_t,$$

where $\Sigma_{t+1} = (\boldsymbol{I} + \eta \sum_{s=1}^{t} \boldsymbol{M}_s)^{-1}$ *and* $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \Sigma_{t+1} \boldsymbol{g}_t$.

EW for Quadratic Lower-bounded Losses

Theorem

Exponential weights with Gaussian prior $P_1 = \mathcal{N}(0, \boldsymbol{I})$ *and constant* $\eta_t = \eta$ *produces Gaussian distributions* $P_{t+1} = \mathcal{N}(\boldsymbol{\theta}_{t+1}, \Sigma_{t+1})$ *and guarantees*

$$\mathrm{Regret}_T(\boldsymbol{\theta}^*) \leq \frac{1}{2\eta}\|\boldsymbol{\theta}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \Sigma_{t+1} \boldsymbol{g}_t,$$

where $\Sigma_{t+1} = (\boldsymbol{I} + \eta \sum_{s=1}^{t} \boldsymbol{M}_s)^{-1}$ *and* $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \Sigma_{t+1} \boldsymbol{g}_t$.

Example 1: Online Regression

- $f_t(\boldsymbol{\theta}) = (Y_t - \boldsymbol{X}_t^\intercal \boldsymbol{\theta})^2, \quad \eta = 1$
- $\boldsymbol{g}_t = 2(\boldsymbol{X}_t^\intercal \boldsymbol{\theta} - Y_t)\boldsymbol{X}_t, \quad \boldsymbol{M}_t = 2\boldsymbol{X}_t \boldsymbol{X}_t^\intercal$

$$\mathrm{Regret}_T(\boldsymbol{\theta}^*) = O\big(\|\boldsymbol{\theta}^*\|^2 + d\log(T)\big)$$

EW for Quadratic Lower-bounded Losses

Theorem

Exponential weights with Gaussian prior $P_1 = \mathcal{N}(0, \boldsymbol{I})$ *and constant* $\eta_t = \eta$ *produces Gaussian distributions* $P_{t+1} = \mathcal{N}(\boldsymbol{\theta}_{t+1}, \Sigma_{t+1})$ *and guarantees*

$$\text{Regret}_T(\boldsymbol{\theta}^*) \leq \frac{1}{2\eta}\|\boldsymbol{\theta}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \Sigma_{t+1} \boldsymbol{g}_t,$$

where $\Sigma_{t+1} = (\boldsymbol{I} + \eta \sum_{s=1}^{t} \boldsymbol{M}_s)^{-1}$ *and* $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \Sigma_{t+1} \boldsymbol{g}_t$.

Example 1: Online Regression

Recovers online ridge regression!

- $f_t(\boldsymbol{\theta}) = (Y_t - \boldsymbol{X}_t^\intercal \boldsymbol{\theta})^2, \quad \eta = 1$
- $\boldsymbol{g}_t = 2(\boldsymbol{X}_t^\intercal \boldsymbol{\theta} - Y_t)\boldsymbol{X}_t, \quad \boldsymbol{M}_t = 2\boldsymbol{X}_t \boldsymbol{X}_t^\intercal$

$$\text{Regret}_T(\boldsymbol{\theta}^*) = O\big(\|\boldsymbol{\theta}^*\|^2 + d\log(T)\big)$$

EW for Quadratic Lower-bounded Losses

Theorem

Exponential weights with Gaussian prior $P_1 = \mathcal{N}(0, \boldsymbol{I})$ and constant $\eta_t = \eta$ produces Gaussian distributions $P_{t+1} = \mathcal{N}(\boldsymbol{\theta}_{t+1}, \Sigma_{t+1})$ and guarantees

$$\mathrm{Regret}_T(\boldsymbol{\theta}^*) \leq \frac{1}{2\eta}\|\boldsymbol{\theta}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \Sigma_{t+1} \boldsymbol{g}_t,$$

where $\Sigma_{t+1} = (\boldsymbol{I} + \eta \sum_{s=1}^{t} \boldsymbol{M}_s)^{-1}$ and $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \Sigma_{t+1} \boldsymbol{g}_t$.

Example 2: Online Logistic Regression

- $f_t(\boldsymbol{\theta}) = \log\left(1 + e^{-Y_t \boldsymbol{X}_t^\intercal \boldsymbol{\theta}}\right)$ for $Y_t \in \{-1, +1\}$
- $\boldsymbol{M}_t = \frac{1+e^{-1}}{4} \boldsymbol{g}_t \boldsymbol{g}_t^\intercal$ if $\|\boldsymbol{X}_t\|_2 \leq 1, \|\boldsymbol{\theta}\|_2 \leq 1$
- $\eta = \frac{1+e^{-1}}{4}$

$$\mathrm{Regret}(\boldsymbol{\theta}^*) = O(d \log T)$$

EW for Quadratic Lower-bounded Losses

Theorem

Exponential weights with Gaussian prior $P_1 = \mathcal{N}(0, \boldsymbol{I})$ and constant $\eta_t = \eta$ produces Gaussian distributions $P_{t+1} = \mathcal{N}(\boldsymbol{\theta}_{t+1}, \Sigma_{t+1})$ and guarantees

$$\text{Regret}_T(\boldsymbol{\theta}^*) \leq \frac{1}{2\eta}\|\boldsymbol{\theta}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \Sigma_{t+1} \boldsymbol{g}_t,$$

where $\Sigma_{t+1} = (\boldsymbol{I} + \eta \sum_{s=1}^{t} \boldsymbol{M}_s)^{-1}$ and $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \Sigma_{t+1} \boldsymbol{g}_t$.

Example 2: Online Logistic Regression

- $f_t(\boldsymbol{\theta}) = \log\left(1 + e^{-Y_t \boldsymbol{X}_t^\intercal \boldsymbol{\theta}}\right)$ for $Y_t \in \{-1, +1\}$
- $\boldsymbol{M}_t = \frac{1+e^{-1}}{4} \boldsymbol{g}_t \boldsymbol{g}_t^\intercal$ if $\|\boldsymbol{X}_t\|_2 \leq 1, \|\boldsymbol{\theta}\|_2 \leq 1$
- $\eta = \frac{1+e^{-1}}{4}$

Recovers online Newton step!

$$\text{Regret}(\boldsymbol{\theta}^*) = O(d \log T)$$

Adaptive Methods for Prediction with Expert Advice

$$\text{Regret}_T(e_i) = O\big(\sqrt{\log(d)T}\big) \qquad \text{for } d \text{ experts}$$

- Q. What if two experts always make the same predictions?
- A. They should count as one expert!

Adaptive Methods for Prediction with Expert Advice

$$\mathrm{Regret}_T(\boldsymbol{e}_i) = O\big(\sqrt{\log(d)T}\big) \qquad \text{for } d \text{ experts}$$

- Q. What if two experts always make the same predictions?
- A. They should count as one expert!

Improvement 1:

$$\sum_{t=1}^{T} f_t(\boldsymbol{\theta}_t) - \mathop{\mathbb{E}}_{Q(i)}\Big[\sum_{t=1}^{T} f_t(\boldsymbol{e}_i)\Big] = O\Big(\sqrt{\mathrm{KL}(Q\|P_1)T}\Big)$$

for all distributions Q on experts

- If $Q = \delta_i$ and P_1 is uniform, then $\mathrm{KL}(Q\|P_1) = \log(d)$.

Adaptive Methods for Prediction with Expert Advice

$$\text{Regret}_T(e_i) = O\big(\sqrt{\log(d)\, T}\big) \qquad \text{for } d \text{ experts}$$

- Q. What if in some round all experts make the same prediction?
- A. Then we should not incur regret on that round!

Adaptive Methods for Prediction with Expert Advice

$$\text{Regret}_T(\boldsymbol{e}_i) = O\big(\sqrt{\log(d)\, T}\big) \qquad \text{for } d \text{ experts}$$

- Q. What if in some round all experts make the same prediction?
- A. Then we should not incur regret on that round!

Improvement 2:

$$\sum_{t=1}^{T} f_t(\boldsymbol{\theta}_t) - \mathop{\mathbb{E}}_{Q(i)}\Big[\sum_{t=1}^{T} f_t(\boldsymbol{e}_i)\Big] = O\Big(\sqrt{\mathrm{KL}(Q\|P_1)\, V_T(Q)}\Big)$$

for all distributions Q on experts,

where $V_T(Q) = \mathbb{E}_{Q(i)}\big[\sum_{t=1}^{T}(f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{e}_i))^2\big] \le T$.

Adaptivity via a Reduction

General Reduction:

- Play distribution $P_t(\eta, i)$ for a surrogate loss $\ell_t(\eta, i)$
- $\boldsymbol{\theta}_t = \frac{\mathbb{E}_{P_t}[\eta \boldsymbol{e}_i]}{\mathbb{E}_{P_t}[\eta]}$

Adaptivity via a Reduction

General Reduction:

- Play distribution $P_t(\eta, i)$ for a surrogate loss $\ell_t(\eta, i)$
- $\boldsymbol{\theta}_t = \dfrac{\mathbb{E}_{P_t}[\eta \boldsymbol{e}_i]}{\mathbb{E}_{P_t}[\eta]}$

Surrogate loss:

- $\ell_t(\eta, i) = -\ln(1 + \eta r_t(i))$ — iProd [Koolen, Van Erven, 2015]
- $\ell_t(\eta, i) = -\eta r_t(i) + \eta^2 r_t(i)^2$ — Squint [Koolen, Van Erven, 2015]
- $\ell_t(\eta, i) = -\frac{1+r_t(i)}{2} \ln(1+\eta) - \frac{1-r_t(i)}{2} \ln(1-\eta)$ — Coin Betting [Orabona, Pál, 2016]

where $r_t(i) := f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{e}_i)$

Adaptivity via a Reduction

General Reduction:

- Play distribution $P_t(\eta, i)$ for a surrogate loss $\ell_t(\eta, i)$
- $\boldsymbol{\theta}_t = \frac{\mathbb{E}_{P_t}[\eta \boldsymbol{e}_i]}{\mathbb{E}_{P_t}[\eta]}$

Surrogate loss:

- $\ell_t(\eta, i) = -\ln(1 + \eta r_t(i))$ iProd [Koolen, Van Erven, 2015]
- $\ell_t(\eta, i) = -\eta r_t(i) + \eta^2 r_t(i)^2$ Squint [Koolen, Van Erven, 2015]
- $\ell_t(\eta, i) = -\frac{1+r_t(i)}{2}\ln(1+\eta) - \frac{1-r_t(i)}{2}\ln(1-\eta)$ Coin Betting [Orabona, Pál, 2016]

where $r_t(i) := f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{e}_i)$

Using Exponential Weights for P_t:

- Coin Betting: improvement 1 $O\left(\sqrt{\mathrm{KL}(Q\|P_1)T}\right)$
- iProd, Squint: improvements 1 and 2 $\tilde{O}\left(\sqrt{\mathrm{KL}(Q\|P_1)V_T(Q)}\right)$

Summary

Setting: Online Convex Optimization for sequential data

- Football prediction and other online regression tasks
- Prediction with expert advice
- Online logistic regression
- ...

The versatile **Exponential Weights** algorithm

By changing the **prior**:

- Optimal for L_1 and L_2-bounded domains
- Recovers mirror descent, online ridge regression, online Newton step
- Recovers Squint and Coin Betting for adaptive prediction with expert advice
- ...

Papers

- D. van der Hoeven, T. van Erven and W. Kotłowski
  The Many Faces of Exponential Weights in Online Learning
  Conference on Learning Theory (COLT), 2018.
- W. M. Koolen and T. van Erven
  Second-order Quantile Methods for Experts and Combinatorial Games
  Conference on Learning Theory (COLT), 2015.