

# 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  $\mathbf{X}_t \in \mathbb{R}^d$ , he may predict  $\hat{Y}_t = \mathbf{X}_t^\top \boldsymbol{\theta}_t$  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**:  $\boldsymbol{\theta}_t \rightarrow \boldsymbol{\theta}_{t+1}$

# 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  $\mathbf{X}_t \in \mathbb{R}^d$ , he may predict  $\hat{Y}_t = \mathbf{X}_t^\top \boldsymbol{\theta}_t$  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**:  $\boldsymbol{\theta}_t \rightarrow \boldsymbol{\theta}_{t+1}$

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

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

# Online Convex Optimization

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

- 1: **for**  $t = 1, 2, \dots, T$  **do**
- 2:   Learner **estimates**  $\theta_t \in \Theta$
- 3:   Nature reveals **convex loss function**  $f_t : \Theta \rightarrow \mathbb{R}$
- 4: **end for**

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

$$V = \min_{\theta_1} \max_{f_1} \min_{\theta_2} \max_{f_2} \cdots \min_{\theta_T} \max_{f_T} \max_{\theta^* \in \Theta} \text{Regret}_T(\theta^*)$$

# Online Gradient Descent

$$\begin{aligned}\tilde{\theta}_{t+1} &= \theta_t - \eta_t \nabla f_t(\theta_t) \\ \theta_{t+1} &= \min_{\theta \in \Theta} \|\tilde{\theta}_{t+1} - \theta\|\end{aligned}$$

## Theorem (Zinkevich, 2003)




Suppose  $\Theta$  compact with diameter at most  $D$ , and  $\|\nabla f_t(\theta_t)\|_2 \leq G$ .  
Then online gradient descent with  $\eta_t = \frac{D}{G\sqrt{t}}$  guarantees

$$\text{Regret}_T(\theta^*) \leq \frac{3}{2} GD\sqrt{T}$$





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	...	Day $T$
Expert 1				...	
Expert 2				...	
Expert 3				...	
Truth				...	

## 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	...	Day $T$
Expert 1				...	
Expert 2				...	
Expert 3				...	
Truth				...	















## 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	...	Day $T$
Expert 1				...	
Expert 2				...	
Expert 3				...	
Truth				...	

















## 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	...	Day $T$
Expert 1				...	
Expert 2				...	
Expert 3				...	
Truth				...	

## 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	...	Day $T$
Expert 1				...	
Expert 2				...	
Expert 3				...	
Truth				...	

### Fits in Framework:

- ▶ Linear loss:  $f_t(\theta) = g_t^\top \theta$   
where  $g_t \in \{0, 1\}^d$  contains mistakes of  $d$  experts
- ▶ Compare with deterministic choice of expert  $\theta^* \in \{e_1, \dots, e_d\}$
- ▶ But allow randomized predictions:  $\theta_t = \mathbb{E}_{P_t(i)}[e_i]$

## Example: Prediction with Expert Advice

	Day 1	Day 2	Day 3	...	Day $T$
Expert 1				...	
Expert 2				...	
Expert 3				...	
Truth				...	

### Fits in Framework:

- ▶ Linear loss:  $f_t(\theta) = g_t^\top \theta$   
where  $g_t \in \{0, 1\}^d$  contains mistakes of  $d$  experts
- ▶ Compare with deterministic choice of expert  $\theta^* \in \{e_1, \dots, e_d\}$
- ▶ But allow randomized predictions:  $\theta_t = \mathbb{E}_{P_t(i)}[e_i]$

**GD Regret Bound**  
 $O(GD\sqrt{T}) = O(\sqrt{dT})$

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

# Exponential Weights for Expert Advice

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

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

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

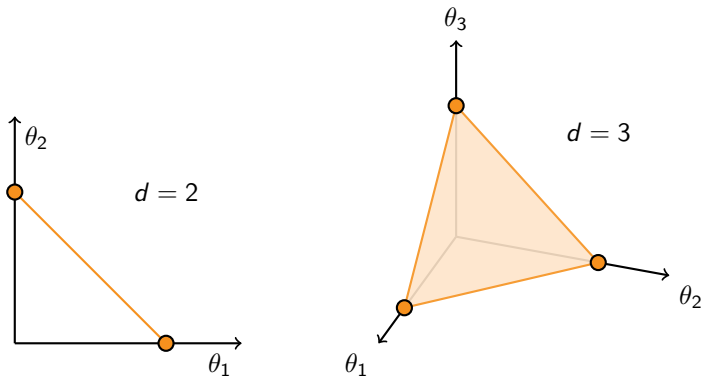
*Exponential weights for expert advice with uniform prior  $P_1$  and  $\eta_t = \sqrt{\frac{8 \log(d)}{T}}$  guarantees*

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

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

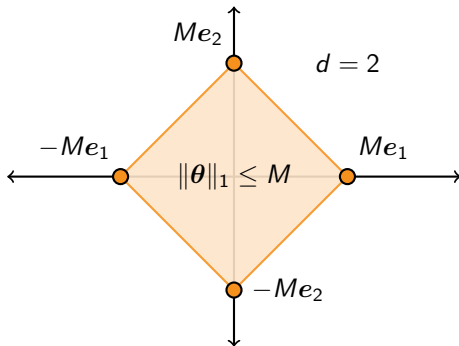
This is optimal for experts (with exactly these constants).

# A Broader View of Exponential Weights



- ▶ Linear loss:  $f_t(\theta) = \mathbf{g}_t^\top \theta$
- ▶ Prior  $P_1$  supported on corners of simplex  $\{e_1, \dots, e_d\}$
- ▶ Distribution of predictions is mean of  $P_t$ :  $\theta_t = \mathbb{E}_{P_t(\theta)}[\theta]$

# Scaling up Exponential Weights



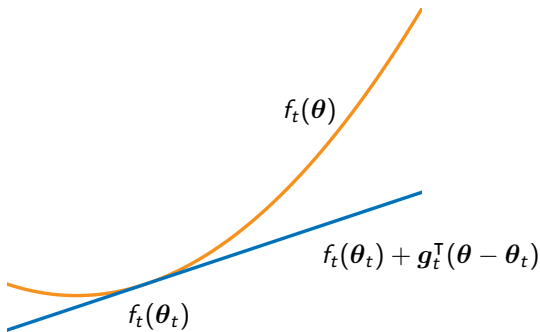
## Theorem (EG<sup>±</sup>, Kivinen, Warmuth, 1997)

Suppose  $f_t(\theta) = g_t^\top \theta$  and  $\|g_t\|_\infty \leq G$ . Then exponential weights with uniform prior on  $\{\pm Me_1, \dots, \pm Me_d\}$  and  $\eta_t = \sqrt{\frac{2 \log(2d)}{M^2 G^2 T}}$  guarantees

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

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

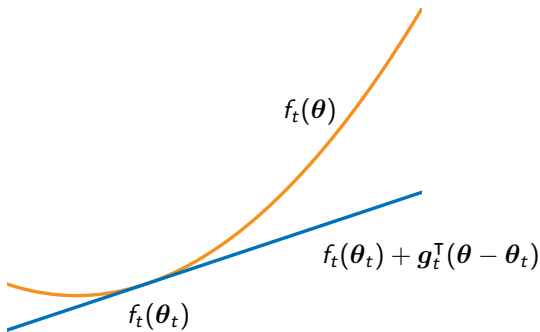
# Linearizing General Convex Losses



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

$$f_t(\theta_t) - f_t(\theta^*) \leq \mathbf{g}_t^T(\theta_t - \theta^*) = \tilde{f}_t(\theta_t) - \tilde{f}_t(\theta^*)$$

# Linearizing General Convex Losses

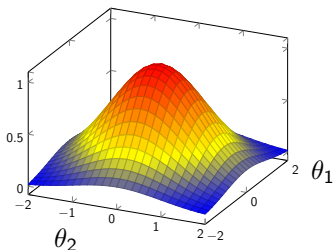


For  $g_t = \nabla f_t(\theta_t)$ , the linear loss  $\tilde{f}_t(\theta) = g_t^T \theta$  satisfies

$$f_t(\theta_t) - f_t(\theta^*) \leq g_t^T(\theta_t - \theta^*) = \tilde{f}_t(\theta_t) - \tilde{f}_t(\theta^*)$$

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

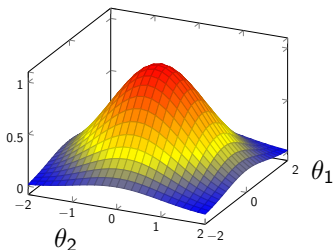
# Exponential Weights for $L_2$ -Domains



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

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

# Exponential Weights for $L_2$ -Domains

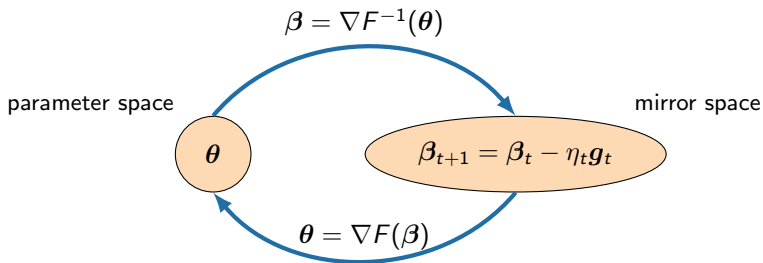


- ▶ **Multivariate Gaussian** prior:  $P_1 = \mathcal{N}(0, \mathbf{I})$
- ▶ **Linearized losses**

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

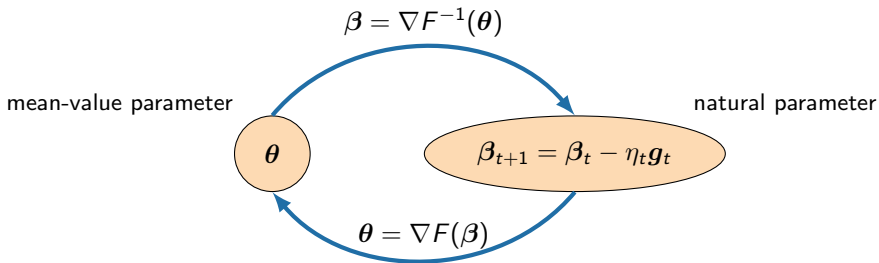
Recover gradient descent!  $P_{t+1} = \mathcal{N}\left(-\eta_t \sum_{s=1}^t \mathbf{g}_s, \mathbf{I}\right)$

# Mirror Descent





# Mirror Descent



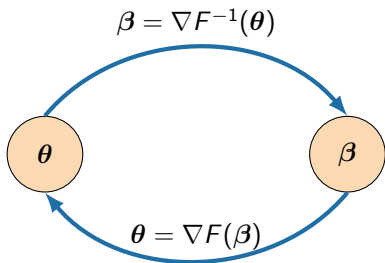
## Theorem

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

$$F(\beta) = \ln \int e^{\beta^\top \theta} dP_1(\theta).$$

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

# Mirror Descent



Examples:

▶ **Gradient descent:**

$$F(\beta) = \frac{1}{2} \|\beta\|_2^2, \quad P_1 = \mathcal{N}(0, I)$$

▶ **Unnormalized relative entropy:**

$$F(\beta) = \sum_{i=1}^d e^{\beta_i},$$

$P_1 = \prod_{i=1}^d P_{\lambda_i}(\theta_i)$  is product of Poisson distributions

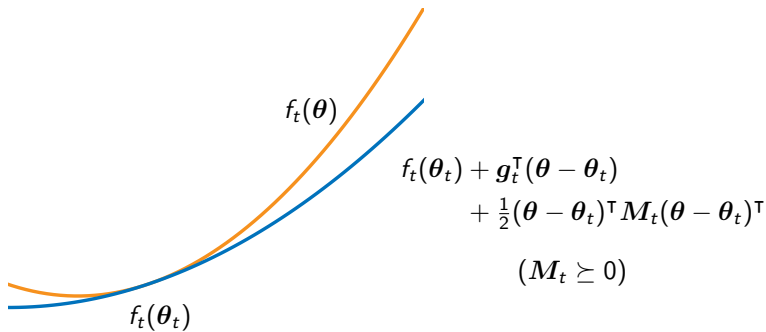
## Theorem

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

$$F(\beta) = \ln \int e^{\beta^\top \theta} dP_1(\theta).$$

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

# Quadratic Lower-bounded Losses



The quadratic loss  $\tilde{f}_t(\boldsymbol{\theta}) = \mathbf{g}_t^\top \boldsymbol{\theta} + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \mathbf{M}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top$  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  $P_1 = \mathcal{N}(0, \mathbf{I})$  and constant  $\eta_t = \eta$  produces Gaussian distributions  $P_{t+1} = \mathcal{N}(\boldsymbol{\theta}_{t+1}, \boldsymbol{\Sigma}_{t+1})$  and guarantees*

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

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

# EW for Quadratic Lower-bounded Losses

## Theorem

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

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

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

## Example 1: Online Regression

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

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

# EW for Quadratic Lower-bounded Losses

## Theorem

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

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

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

## Example 1: Online Regression

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

Recovers online ridge regression!

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

# EW for Quadratic Lower-bounded Losses

## Theorem

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

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

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

## Example 2: Online Logistic Regression

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

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

# EW for Quadratic Lower-bounded Losses

## Theorem

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

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

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

## Example 2: Online Logistic Regression

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

Recovers online Newton step!

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



# Adaptive Methods for Prediction with Expert Advice

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

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

# Adaptive Methods for Prediction with Expert Advice

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

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

## Improvement 1:

$$\sum_{t=1}^T f_t(\theta_t) - \mathbb{E}_{Q(i)} \left[ \sum_{t=1}^T f_t(e_i) \right] = O\left(\sqrt{\text{KL}(Q\|P_1)T}\right)$$

for all distributions  $Q$  on experts

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

# Adaptive Methods for Prediction with Expert Advice

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

- ▶ 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

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

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

## Improvement 2:

$$\sum_{t=1}^T f_t(\theta_t) - \mathbb{E}_{Q(i)} \left[ \sum_{t=1}^T f_t(e_i) \right] = O\left(\sqrt{\text{KL}(Q\|P_1) V_T(Q)}\right)$$

for all distributions  $Q$  on experts,

where  $V_T(Q) = \mathbb{E}_{Q(i)} \left[ \sum_{t=1}^T (f_t(\theta_t) - f_t(e_i))^2 \right] \leq T$ .

# Adaptivity via a Reduction

## General Reduction:

- ▶ Play distribution  $P_t(\eta, i)$  for a surrogate loss  $\ell_t(\eta, i)$
- ▶  $\theta_t = \frac{\mathbb{E}_{P_t}[\eta e_i]}{\mathbb{E}_{P_t}[\eta]}$

# Adaptivity via a Reduction

## General Reduction:

- ▶ Play distribution  $P_t(\eta, i)$  for a surrogate loss  $\ell_t(\eta, i)$
- ▶  $\theta_t = \frac{\mathbb{E}_{P_t}[\eta e_i]}{\mathbb{E}_{P_t}[\eta]}$

## 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  $r_t(i) := f_t(\theta_t) - f_t(e_i)$

# Adaptivity via a Reduction

## General Reduction:

- ▶ Play distribution  $P_t(\eta, i)$  for a surrogate loss  $\ell_t(\eta, i)$
- ▶  $\theta_t = \frac{\mathbb{E}_{P_t}[\eta e_i]}{\mathbb{E}_{P_t}[\eta]}$

## 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  $r_t(i) := f_t(\theta_t) - f_t(e_i)$

## Using Exponential Weights for $P_t$ :

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

# 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.