Reminder

- Quiz for Lecture 4 (9/15, 10pm)
- Paper Bidding Result
  - Next Mon's presenter
- Paper Reading Assignment 1 (9/13, 10pm)
  - Peer reviewed (Due 1 week after assignment due)
- Confirm group members for course project (9/13, 10pm)

Advanced Topics in
Machine Learning and Game Theory
Lecture 5: Introduction to Online Learning

17599/17759
Fei Fang
feifang@cmu.edu

Outline

- Online Learning
- Regret Analysis
- Follow-the-(Regularized)-Leader
- Online Mirror Descent

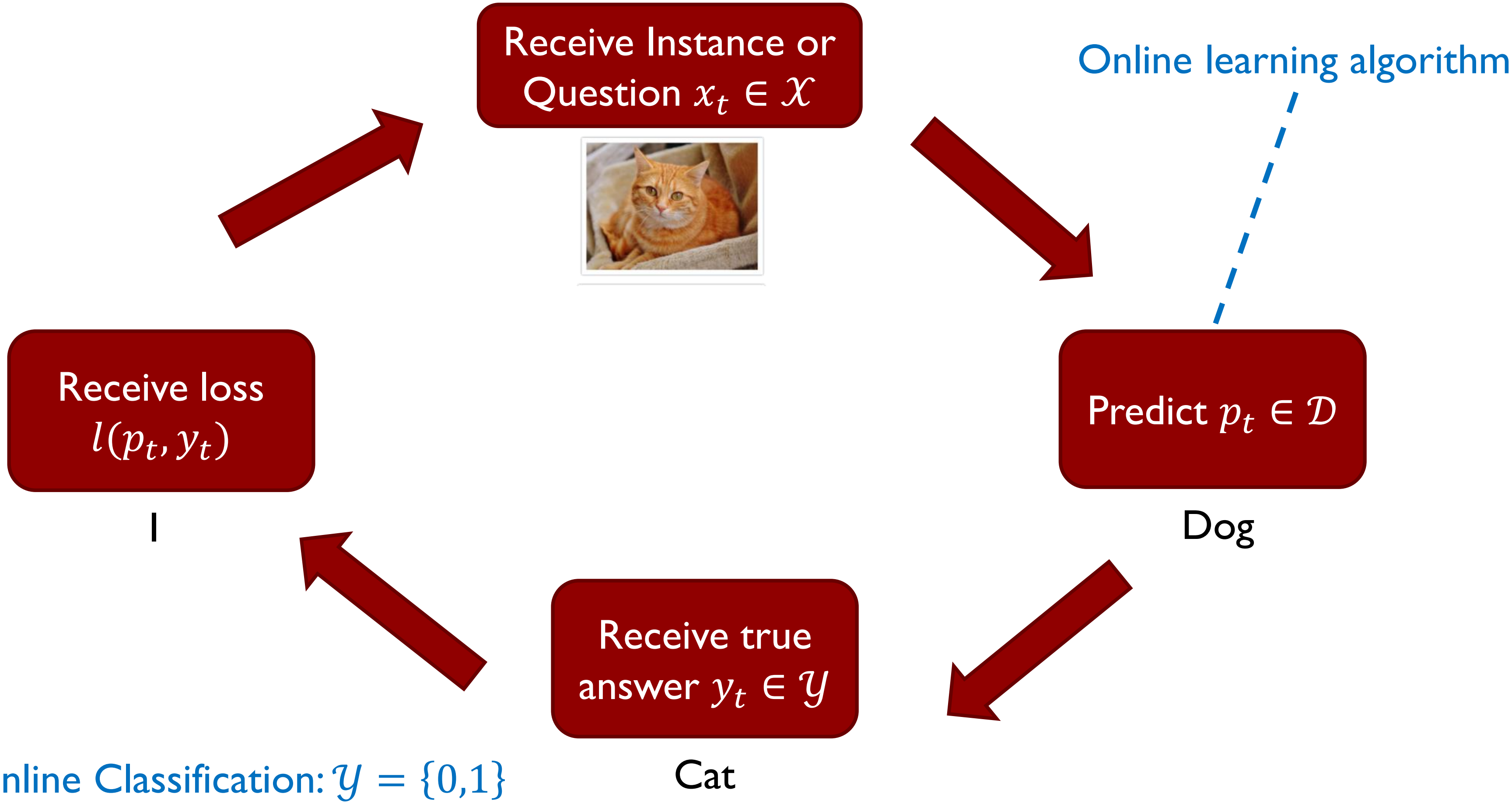
Online Learning

- Supervised Learning: Learn from a dataset with labels
- Unsupervised Learning: Learn from a dataset without labels
- Online Learning: Data come online

Cat or Dog?

Chihuahua or Muffin?

Online Learning Pipeline



Online Classification: $\mathcal{Y} = \{0,1\}$
$\mathcal{D} = \{0,1\}$ or $[0,1]$
$l_t(p_t, y_t) = |p_t - y_t|$

Full feedback: Observe y_t and $l(p_t, y_t)$
Limited feedback (bandit): Only observe $l(p_t, y_t)$

Online Learning Pipeline

Receive Instance or Question $x_t \in \mathcal{X}$

meteorological measurements

Predict $p_t \in \mathcal{D}$

Rain tomorrow

Receive true answer $y_t \in \mathcal{Y}$

No rain

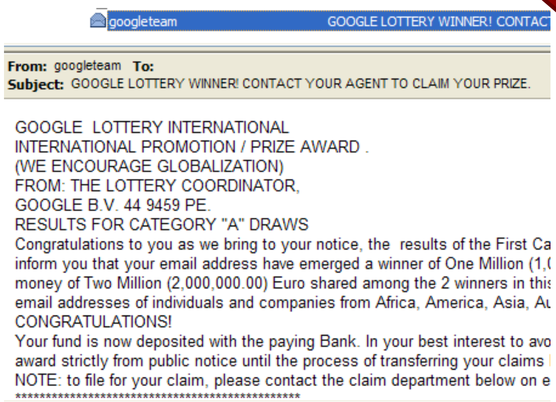
Receive loss $l(p_t, y_t)$

1

https://www.ecotech-bonn.de/en/produkte/meterologie/anwendungen_projekte3/messturm_autarke_messstation_1/

Online Learning Pipeline

Receive Instance or Question $x_t \in \mathcal{X}$


googleteam GOOGLE LOTTERY WINNER! CONTACT

From: googleteam To:
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,(
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****

Predict $p_t \in \mathcal{D}$

Spam

Receive true answer $y_t \in \mathcal{Y}$

Spam

Receive loss $l(p_t, y_t)$

0

The spam designer may adapt to learner's learning algorithm!

https://group-mail.com/wp-content/uploads/what-is-spam-scam.gif

Online Learning Pipeline

Receive Instance or Question $x_t \in \mathcal{X}$

Experts' Prediction on Tomorrow's Stock Price

Expt 1	Expt 2	Expt 3	Expt 4
Up	Down	Down	Down

Predict $p_t \in \mathcal{D}$

Follow Expt 3's Advice, i.e., Down

Receive true answer $y_t \in \mathcal{Y}$

Up

Receive loss $l(p_t, y_t)$

1

Prediction with Expert Advice

Online Learning Pipeline

Online Ranking

Online Learning Pipeline

Receive Instance or Question $x_t \in \mathcal{X}$

Lot size, #br, #ba

Predict $p_t \in \mathcal{D}$

House Selling Price

Receive true answer $y_t \in \mathcal{Y}$

Actual Selling Price

Receive loss $l(p_t, y_t)$

Squared price difference

If we assume the actual selling price is a linear function of the features: Online Regression

https://www.zillow.com/research/strategy-best-time-to-buy-15066/

Stochastic vs Adversarial Online Learning

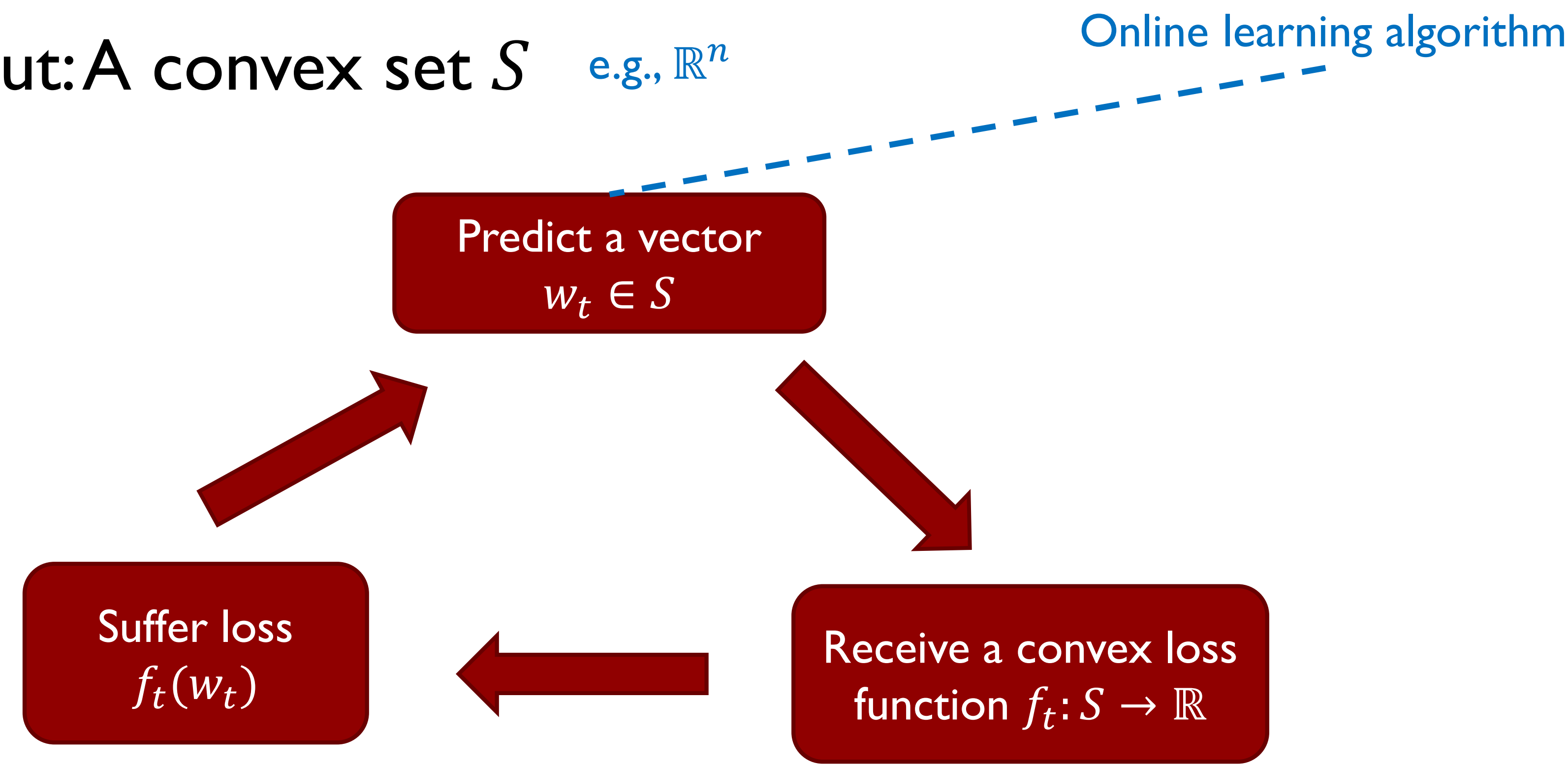
- Stochastic/statistical setting: instances are drawn i.i.d. from a fixed distribution
  - Image classification, predict stock prices
- Adversarial setting: an adversary picks the worst instance at every time step (adapt to learner's past actions and even the learner's learning algorithm)
  - Spam detection, anomaly detection, game playing

Applications of Online Learning

- Learn to make decisions in daily life
  - How to commute to school? Bus, walking, or driving? Which route?
- Learn to gamble or buy stocks
- Advertisers learn to bid for keywords
- Others?

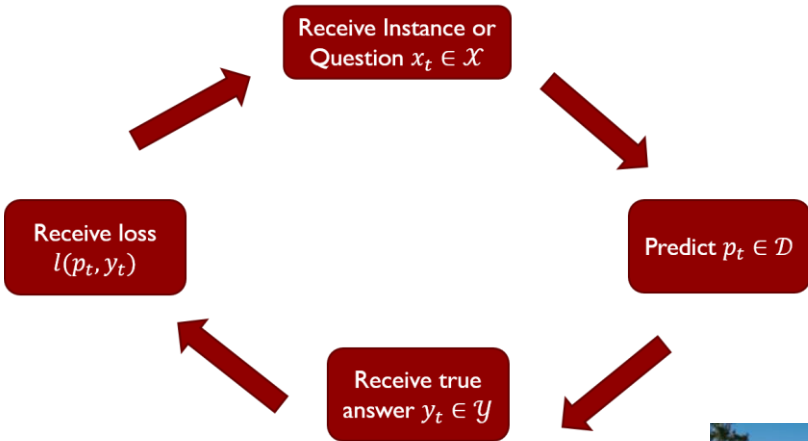
Online Convex Optimization

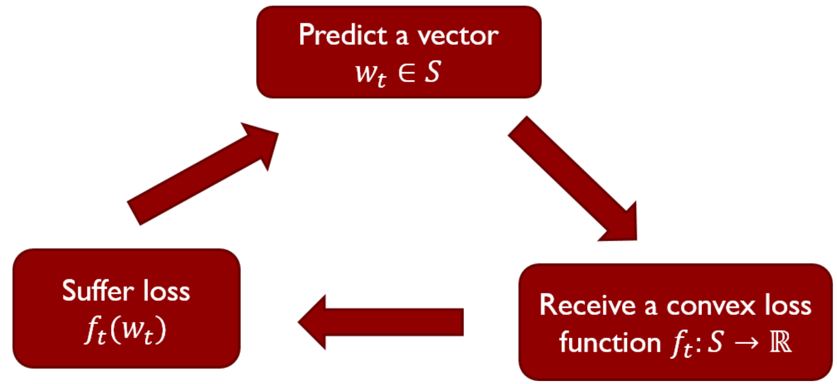
- A more abstract model
- Input: A convex set S e.g., $\mathbb{R}^n$



Online Convex Optimization

Convexity is preserved under a linear transformation: If $f(x) = g(Ax + b)$, g convex, then $f(x)$ is convex



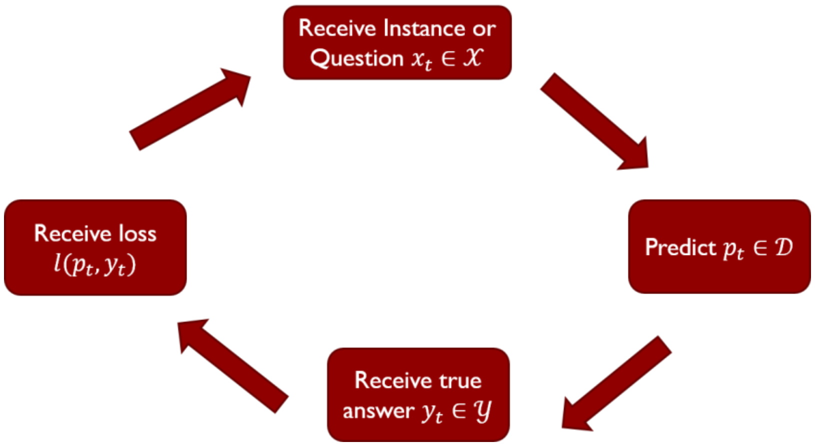


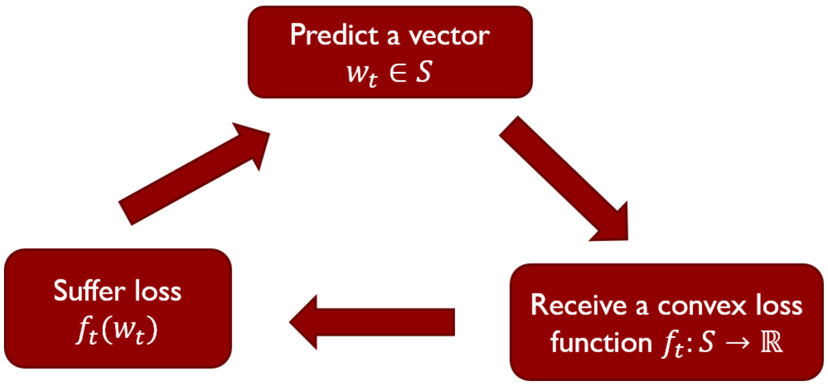
Online Regression: w_t are the parameters in the linear regression model

$$p_t = \sum_i w_t[i] x_t[i] = \langle w_t, x_t \rangle$$

$$f_t(w_t) = l(p_t, y_t) = \left(\sum_i w_t[i] x_t[i] - y_t \right)^2$$

Online Convex Optimization





Expt 1	Expt 2	Expt 3	Expt 4
Up	Down	Down	Down

Prediction with Expert Advice: If there are n experts

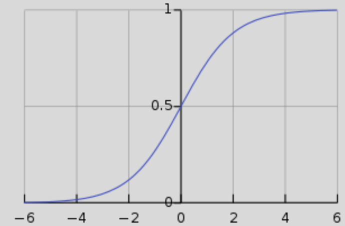
$w_t \in \mathbb{R}^n$ are the probabilities of following each expert's advice

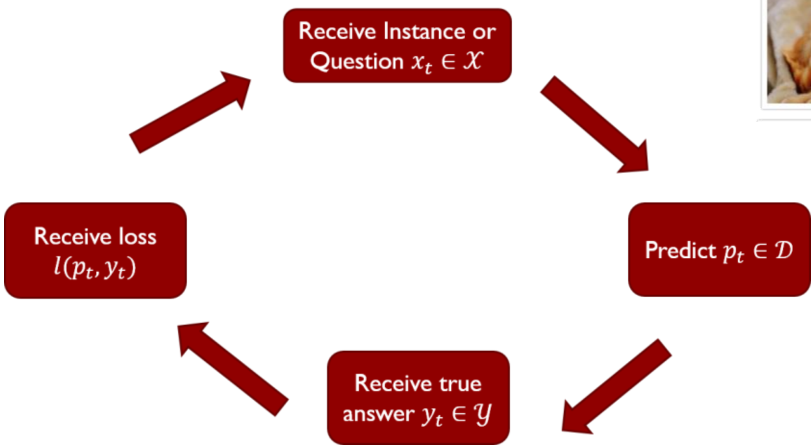
$p_t \sim w_t$, i.e., $\mathbb{P}[p_t = i] = w_t[i]$

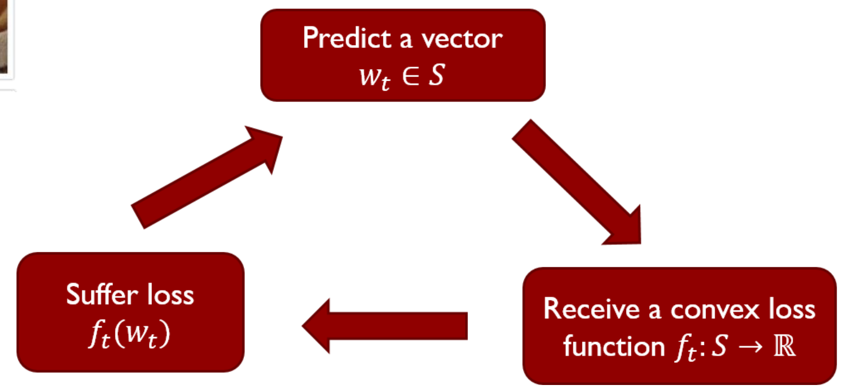
$f_t(w_t) = \mathbb{E}_{p_t \sim w_t}[l(p_t, y_t)]$

Quiz 1

$$sigmoid(a) = \frac{1}{1 + e^{-a}}$$







Assume we use a simple model for online image classification:

$$p_t = g\left(\sum_i w_t[i] x_t[i]\right)$$

g maps the linear combination to $[0,1]$, e.g., sigmoid

When can the online image classification problem be described as an OCO problem?

A: $l(p_t, y_t)$ is a convex function of p_t

B: $f_t(w_t)$ is a convex function of w_t

C: $g(a)$ is a convex function of a

Outline

- Online Learning
- Regret Analysis
- Follow-the-(Regularized)-Leader
- Online Mirror Descent

Regret

- How "sorry" the learner is in retrospect
- In online classification
  - $\mathcal{Y} = \{0,1\}, \mathcal{D} = \{0,1\}$ or $[0,1]$ (randomize over $\{0,1\}$)
  - $l_t(p_t, y_t) = |p_t - y_t|$
  - An online learning algorithm A makes predictions p_t
  - After T time steps, regret relative to a fixed predictor $h^\star: \mathcal{X} \to \mathcal{Y} = \{0,1\}$ is

$$\mathrm{Regret}_T(h^\star) = \sum_{t=1}^{T} l(p_t, y_t) - \sum_{t=1}^{T} l(h^\star(x_t), y_t)$$

  - Regret relative to a hypothesis class $\mathcal{H}$ is

$$\mathrm{Regret}_T(\mathcal{H}) = \max_{h^\star \in \mathcal{H}} \mathrm{Regret}_T(h^\star)$$

Compare to the best fixed hypothesis in hindsight

Regret

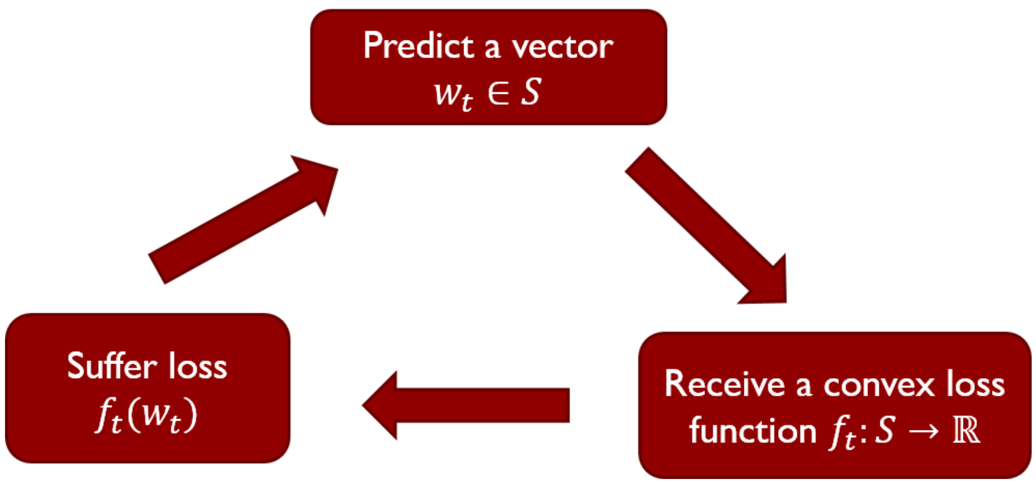
- Generally, in online convex optimization
- Regret w.r.t. some vector u is

$$\text{Regret}_T(u) = \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u)$$

- Regret w.r.t. a set of vectors U is

$$\text{Regret}_T(U) = \max_{u \in U} \text{Regret}_T(u)$$

Compare to the best fixed vector in U in hindsight



No Regret

- Consider the average regret $\bar{R} = \frac{\text{Regret}_T}{T}$
- If $\bar{R} \to 0$ as $T \to \infty$, we say the online learning algorithm has no-regret
  - Equivalently, we can say, the regret is sublinear in T
- A typical goal in online learning is to design no-regret algorithms

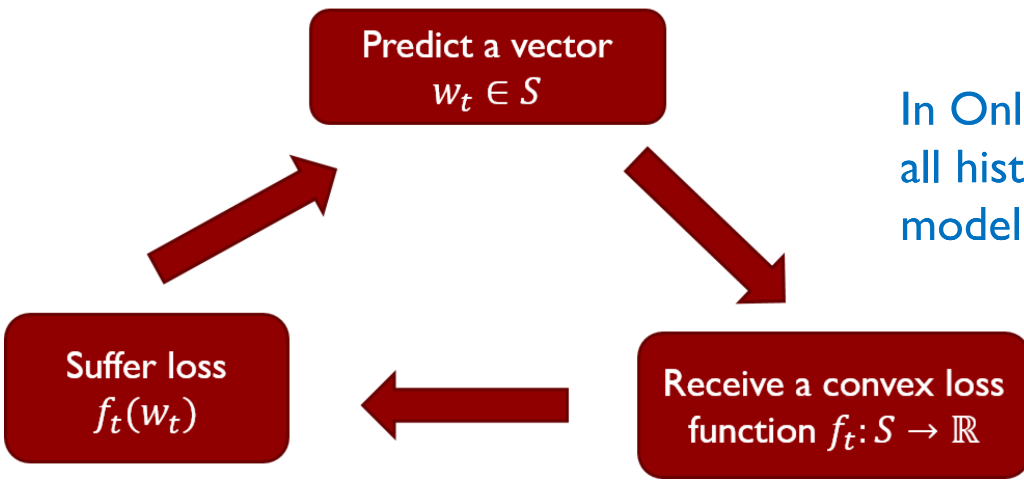
Outline

- Online Learning
- Regret Analysis
- Follow-the-(Regularized)-Leader
- Online Mirror Descent

Follow-the-Leader (FTL)

Follow-the-Leader
$$\forall t, w_t = \text{argmin}_{w \in S} \sum_{i}^{t-1} f_i(w)$$

- Pick the best vector on all past rounds
- Break ties arbitrarily



In Online Regression: Train a model with all historical data, and use the trained model for prediction in the next round

Quiz 2

- If we apply FTL to Prediction with Expert Advice, which expert's advice will be followed in each round? (Assume the expert's advice is binary)
  - A: Probability of choosing expert i is proportional to the number of past rounds expert i is correct
  - B: Always follow the expert with the minimum number of mistakes in the past rounds
  - C: None of the above

Follow-the-Leader

$$\forall t, w_t = \text{argmin}_{w \in S} \sum_{i}^{t-1} f_i(w)$$

Prediction with Expert Advice: If there are n experts

$w_t \in \mathbb{R}^n$ are the probabilities of following each expert's advice

$p_t \sim w_t$, i.e., $\mathbb{P}[p_t = i] = w_t[i]$

$f_t(w_t) = \mathbb{E}_{p_t \sim w_t}[l(p_t, y_t)]$

Follow-the-Regularized-Leader (FoReL)

Follow-the-Regularizaed-Leader

$$\forall t, w_t = \text{argmin}_{w \in S} \sum_{i}^{t-1} f_i(w) + R(w)$$

- Use a regularization function
- Different regularization functions will yield different algorithms with different regret bounds

FoReL

$\forall t, w_t = \text{argmin}_{w \in S} \sum_{i}^{t-1} f_i(w) + R(w)$

- Consider a problem where $f_t(w) = \langle w, z_t \rangle$ for some vector z_t and $S = \mathbb{R}^d$
- Run FoReL with regularization function $R(w) = \frac{1}{2\eta} ||w||_2^2$ for some positive scalar η
- Then $w_{t+1} =$

Online gradient descent!

FoReL

$$\forall t, w_t = \text{argmin}_{w \in S} \sum_i^{t-1} f_i(w) + R(w)$$

- Consider a problem where $f_t(w) = \langle w, z_t \rangle$ for some vector z_t and $S = \mathbb{R}^d$
- Run FoReL with regularization function $R(w) = \frac{1}{2\eta} \|w\|_2^2$ for some positive scalar η
- Then $w_{t+1} = \text{argmin}_{w \in S} \sum_i^t f_i(w) + R(w) = \text{argmin}_{w \in S} \sum_i^t w^T z_t + \frac{1}{2\eta} \|w\|_2^2$

Set gradient of the function w.r.t w to be 0 to get w_{t+1}, i.e.,

$$\sum_i^t z_t + \frac{1}{2\eta} 2w = 0$$

So $w_{t+1} = -\eta \sum_{i=1}^t z_t = w_t - \eta z_t = w_t - \eta \partial f_t(w_t)$

Online gradient descent!

FoReL

- It can be proved that running this version of FoReL on this problem yield

$$\text{Regret}_T(u) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\|z_t\|_2^2, \forall u$$

- If we consider a set of vectors $U = \{u: \|u\| \leq B\}$, with a properly chosen constant η, we can get

$$\text{Regret}_T(U) \leq BL\sqrt{2T}$$

Is this version of FoReL a no-regret algorithm for the problem?

Disadvantage of FoReL

- Need to solve an optimization problem at each online round

FoReL

$$\forall t, w_t = \text{argmin}_{w \in S} \sum_{i}^{t-1} f_i(w) + R(w)$$

Outline

- Online Learning
- Regret Analysis
- Follow-the-(Regularized)-Leader
- Online Mirror Descent

Online Mirror Descent (OMD)

- A family of algorithms without solving an optimization problem in each round

Online Mirror Descent
Parameters: a link function $g: \mathbb{R}^d \to S$
Initialize: $\theta_1 = 0$
for $t = 1,2, \ldots$
predict $w_t = g(\theta_t)$
Update $\theta_{t+1} = \theta_t - z_t$ where $z_t = \partial f_t(w_t)$

- Different link functions will yield different algorithms with different regret bounds

Quiz 3

- If $S = \mathbb{R}^d, g(\theta) = \eta\theta$, what is the relationship between w_{t+1} and w_t?
  - A: $w_{t+1} \geq w_t$
  - B: $w_{t+1} \leq w_t$
  - C: $w_{t+1} = w_t - \eta\partial f_t(w_t)$
  - D: $w_{t+1} = w_t - \eta\theta_t$
  - E: None of the above

Online Mirror Descent

Parameters: a link function $g: \mathbb{R}^d \rightarrow S$
Initialize: $\theta_1 = 0$
for $t = 1,2, \ldots$
    predict $w_t = g(\theta_t)$
    Update $\theta_{t+1} = \theta_t - z_t$ where $z_t = \partial f_t(w_t)$

Quiz 3

- If $S = \mathbb{R}^d, g(\theta) = \eta\theta$, what is the relationship between w_{t+1} and w_t? Online gradient descent again!
  - A: $w_{t+1} \geq w_t$
  - B: $w_{t+1} \leq w_t$
  - C: $w_{t+1} = w_t - \eta\partial f_t(w_t)$
  - D: $w_{t+1} = w_t - \eta\theta_t$
  - E: None of the above

Online Mirror Descent

Parameters: a link function $g: \mathbb{R}^d \rightarrow S$
Initialize: $\theta_1 = 0$
for $t = 1,2, \ldots$
predict $w_t = g(\theta_t)$
Update $\theta_{t+1} = \theta_t - z_t$ where $z_t = \partial f_t(w_t)$

Discussion

- Suppose we are playing a two-player normal-form game repeatedly. Can this be described as an online learning problem? An online convex optimization problem? What would FTL and FoReL mean?

Additional Resources

- Online Learning and Online Convex Optimization, Chp 1-3

Acknowledgment

- The slides are prepared based on slides made by Haifeng Xu

Backup Slides

Multi-Armed Bandit (MAB)

- K arms
- Each arm k is associated with a reward distribution R_k (pdf $p_k(r)$), with expected reward μ_k ($\mu_k = \int_r rp_k(r)dr$)
- Gambler does not know R_k, μ_k
- In each round $t \in \{1 \dots T\}$, gambler chooses one arm k_t, and observe a reward $\hat{r}_t$ drawn from the distribution
- Task: design an online learning algorithm A
- Example Goal: find the best arm with a minimum number of arm pulls

Stochastic feedback
Limited feedback

Regret

- Let $\mu^* = \max_k \mu_k$
- Regret $\rho = T\mu^* - \sum_{t=1}^{T} \widehat{r_t}$
- A typical research problem in MAB: find zero-regret strategy
  - $\lim_{T\to\infty} \frac{\rho}{T} = 0$
- Probably approximately correct (PAC): with high probability, it is close to being correct

$$\Pr(error \leq \epsilon) \geq 1 - \delta$$

- PAC version of zero-regret strategy

$$\Pr(\lim_{T\to\infty} \frac{\rho}{T} \leq \epsilon) \geq 1 - \delta$$

Binary MAB

- K arms
- Reward is either 0 or 1, R_k: $\Pr(r = 1) = p_k, \Pr(r =$

Upper Confidence Bound in Binary MAB

- Let $N(k)$ be the number of times that k is chosen
- Let $H(k)$ be the number of times that k is chosen and reward is 1
- Let $\widehat{\mu_k} = H(k)/N(k)$, average reward when k is chosen
- Given $N(k), H(k), \widehat{\mu_k}, \delta$, we can estimate the range of μ_k, i.e., we can compute μ_{LB}^k and μ_{UB}^k such that $\Pr\left(\mu_{LB}^k \leq \mu_k \leq \mu_{UB}^k\right) \geq 1 - \delta$

Upper Confidence Bound in Binary MAB

- Chernoff-Hoeffding Bound: Let $X_1, X_2, \ldots, X_n$ be independent random variables in the range $[0,1]$ with $\mathbb{E}[X_i] = \mu$. Then for $a > 0$

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + a\right) \leq e^{-2a^2 n}$$

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq \mu - a\right) \leq e^{-2a^2 n}$$

- That is, with high probability, the observed average value of X_i is very close to the expected value of X_i

Upper Confidence Bound in Binary MAB

- $\widehat{\mu_k} = H(k)/N(k)$
- According to Chernoff-Hoeffding Bound
- $\Pr(\widehat{\mu_k} \geq \mu_k + a) \leq e^{-2a^2 N(k)}$
- $\Pr(\widehat{\mu_k} \leq \mu_k - a) \leq e^{-2a^2 N(k)}$
- So $\Pr(\widehat{\mu_k} - a \leq \mu_k \leq \widehat{\mu_k} + a) \leq 1 - 2e^{-2a^2 N(k)}$

- Given δ, if we want to find μ_{LB}^k and μ_{UB}^k such that $\Pr\left(\mu_{LB}^k \leq \mu_k \leq \mu_{UB}^k\right) \geq 1 - \delta$, then a simple way is to set $\delta = 2e^{-2a^2 N(k)}$, i.e., a $= \sqrt{\frac{1}{2N(k)} \ln(\frac{2}{\delta})}$ and
- $\mu_{LB}^k = \widehat{\mu_k} - a, \mu_{UB}^k = \widehat{\mu_k} + a$

Upper Confidence Bound in Binary MAB

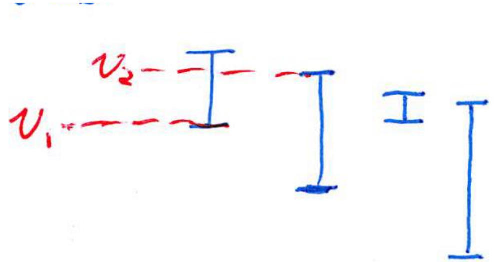
- Heuristic strategy in binary MAB with the goal of finding an arm k such that $\Pr[\mu^* - \mu_k \le \epsilon] \ge 1 - \delta$ with minimum number of arm pulls (rounds)
  - In very round, choose the arm with highest μ_{UB}^k. Terminates when $\mu_{UB}^k - \mu_{LB}^k \le \epsilon$ for the chosen arm.
  - Intuition: If μ_{UB}^k is large, either k is a good arm or $N(k)$ is small (not enough data is gathered)

Upper Confidence Bound in Binary MAB

- Q: When the confidence interval of the arm with highest upper bound is smaller than ϵ, then is the difference between the optimal value and the average value of this arm guaranteed to be smaller than ϵ?

Upper Confidence Bound in Binary MAB

- Q: When the confidence interval of the arm with highest upper bound is smaller than ϵ, then is the difference between the optimal value and the average value of this arm guaranteed to be smaller than ϵ?



$v_2 - v_1 \leq \epsilon$

$\mu_k \geq \mu_{UB}^{k} - \epsilon$

$\mu_{k'} \leq \mu_{UB}^{k'} \leq \mu_{UB}^{k}$

So $\mu_{k'} - \mu_{k'} \leq \epsilon$

Upper Confidence Bound in Binary MAB

- Heuristic strategy in binary MAB with the goal of maximizing accumulated reward: in very round, choose the arm with highest $\mu_{UB}^k = \widehat{\mu_k} + \sqrt{\frac{2\ln(N)}{N(k)}}$

Previously, to ensure
$\Pr\left(\mu_{LB}^k \leq \mu_k \leq \mu_{UB}^k\right) \geq 1 - \delta$
We set $\mu_{UB}^k = \widehat{\mu_k} + a$

$$a = \sqrt{\frac{1}{2N(k)}\ln\left(\frac{2}{\delta}\right)}$$

Upper Confidence Bound

- Extend UCB to general MDP/RL setting
  - Recall in Q-Learning and SARSA, we need to follow some policy (based on current estimates of Q-value)
  - At state s, choose action a with the highest $Q_{UB}(s, a)$
    - $Q_{UB}(s, a) = Q(s, a) + c\sqrt{\frac{\ln N(s)}{N(s,a)}}$
  - Better than ϵ-Greedy in handling exploitation vs exploration tradeoff