Online Prediction

Peter Bartlett

Statistics and EECS
UC Berkeley

and

Mathematical Sciences
Queensland University of Technology

Online Prediction

- Repeated game:

$$\text{Decision method plays } a_t \in \mathcal{A}$$
$$\text{World reveals } \ell_t \in \mathcal{L}$$

- Cumulative loss: $\hat{L}_n = \sum_{t=1}^{n} \ell_t(a_t)$.
- Aim to minimize regret, that is, perform well compared to the best (in retrospect) from some class:

$$\text{regret} = \underbrace{\sum_{t=1}^{n} \ell_t(a_t)}_{\hat{L}_n} - \underbrace{\min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a)}_{L_n^*}.$$

- Data can be adversarially chosen.

Online Prediction

Minimax regret is the value of the game:

$$\min_{a_1} \max_{\ell_1} \cdots \min_{a_n} \max_{\ell_n} \left(\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right).$$

Online Prediction: Motivations

1. Adversarial model is often appropriate, e.g., in
   - Computer security.
   - Computational finance.
2. Adversarial model assumes little:
   It is often straightforward to convert a strategy for an adversarial environment to a method for a probabilistic environment.
3. Studying the adversarial model can reveal the *deterministic core* of a statistical problem: there are strong similarities between the performance guarantees in the two cases.
4. There are significant overlaps in the design of methods for the two problems:
   - *Regularization* plays a central role.
   - Many online prediction strategies have a natural interpretation as a *Bayesian method*.

Computer Security: Spam Detection

Computer Security: Spam Email Detection

- Here, the action a_t might be a classification rule, and ℓ_t is the indicator for a particular email being incorrectly classified (e.g., spam allowed through).
- The sender can determine if an email is delivered (or detected as spam), and try to modify it.
- An adversarial model allows an arbitrary sequence.
- We cannot hope for good classification accuracy in an absolute sense; regret is relative to a comparison class.
- Minimizing regret ensures that the spam detection accuracy is close to the best performance in retrospect on the particular email sequence.

Computer Security: Spam Email Detection

- Suppose we consider features of email messages from some set $\mathcal{X}$ (e.g., information about the header, about words in the message, about attachments).
- The decision method's action a_t is a mapping from $\mathcal{X}$ to $[0, 1]$ (think of the value as an estimated probability that the message is spam).
- At each round, the adversary chooses a feature vector $x_t \in \mathcal{X}$ and a label $y_t \in \{0, 1\}$, and the loss is defined as

$$\ell_t(a_t) = (y_t - a_t(x_t))^2 .$$

- The regret is then the excess squared error, over the best achievable on the data sequence:

$$\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) = \sum_{t=1}^{n} (y_t - a_t(x_t))^2 - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} (y_t - a(x_t))^2 .$$

Computational Finance: Portfolio Optimization

Computational Finance: Portfolio Optimization

- Aim to choose a portfolio (distribution over financial instruments) to maximize utility.
- Other market players can profit from making our decisions bad ones. For example, if our trades have a market impact, someone can *front-run* (trade ahead of us).
- Here, the action a_t is a distribution on instruments, and ℓ_t might be the negative logarithm of the portfolio's increase, $a_t \cdot r_t$, where r_t is the vector of relative price increases.
- We might compare our performance to the best stock (distribution is a delta function), or a set of indices (distribution corresponds to Dow Jones Industrial Average, etc), or the set of all distributions.

Computational Finance: Portfolio Optimization

- The decision method's action a_t is a distribution on the m instruments, $a_t \in \Delta^m = \{a \in [0,1]^m : \sum_i a_i = 1\}$.
- At each round, the adversary chooses a vector of returns $r_t \in \mathbb{R}_+^m$; the ith component is the ratio of the price of instrument i at time t to its price at the previous time, and the loss is defined as

$$\ell_t(a_t) = -\log(a_t \cdot r_t).$$

- The regret is then the log of the ratio of the maximum value the portfolio would have at the end (for the best mixture choice) to the final portfolio value:

$$\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) = \max_{a \in \mathcal{A}} \sum_{t=1}^{n} \log(a \cdot r_t) - \sum_{t=1}^{n} \log(a_t \cdot r_t).$$

Online Prediction: Motivations

2. Online algorithms are also effective in probabilistic settings.

- Easy to convert an online algorithm to a batch algorithm.
- Easy to show that good online performance implies good i.i.d. performance, for example.

Online Prediction: Motivations

3. Understanding statistical prediction methods.

- Many statistical methods, based on *probabilistic assumptions*, can be effective in an adversarial setting.
- Analyzing their performance in adversarial settings provides perspective on their robustness.
- We would like violations of the probabilistic assumptions to have a limited impact.

Key Points

- Online Prediction:
  - repeated game.
  - aim to minimize *regret*.
  - Data can be *adversarially* chosen.
- Motivations:
  - Often appropriate (security, finance).
  - Algorithms also effective in probabilistic settings.
  - Can provide insight into statistical prediction methods.

Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
  - An easy start.
- Online, adversarial versus batch, probabilistic.
  - Similar bounds.
- Optimal regret: dual game.
  - Rademacher averages for probabilistic.
  - Sequential Rademacher averages for adversarial.
- Online convex optimization.
  - Regularization methods.

Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
- Online, adversarial versus batch, probabilistic.
- Optimal regret.
- Online convex optimization.

Finite Comparison Class

1. "Prediction with expert advice."
2. With perfect predictions: $\log m$ regret.
3. Exponential weights strategy: $\sqrt{n \log m}$ regret.
4. Refinements and extensions:
   - Exponential weights and $L^* = 0$
   - n unknown
   - L^* unknown
   - *Convex* (versus linear) losses
   - Bayesian interpretation
5. Probabilistic prediction with a finite class.

Prediction with Expert Advice

Suppose we are predicting whether it will rain tomorrow. We have access to a set of m experts, who each make a forecast of 0 or 1. Can we ensure that we predict almost as well as the best expert?

Here, $\mathcal{A} = \{1, \ldots, m\}$. There are m experts, and each has a forecast sequence $f_1^i, f_2^i, \ldots$ from $\{0, 1\}$. At round t, the adversary chooses an outcome $y_t \in \{0, 1\}$, and sets

$$\ell_t(i) = 1[f_t^i \neq y_t] = \begin{cases} 1 & \text{if } f_t^i \neq y_t, \\ 0 & \text{otherwise.} \end{cases}$$

Online Prediction

Minimax regret is the value of the game:

$$\min_{a_1} \max_{\ell_1} \cdots \min_{a_n} \max_{\ell_n} \left(\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right).$$

$$\hat{L}_n = \sum_{t=1}^{n} \ell_t(a_t), \qquad L_n^* = \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a).$$

Prediction with Expert Advice

An easier game: suppose that the adversary is constrained to choose the sequence y_t so that some expert incurs no loss ($L_n^* = 0$), that is, there is an $i^* \in \{1, \ldots, m\}$ such that for all t, $y_t = f_t^{i^*}$.
How should we predict?

Prediction with Expert Advice: Halving

- Define the set of experts who have been correct so far:

$$C_t = \{i : \ell_1(i) = \cdots = \ell_{t-1}(i) = 0\}.$$

- Choose a_t any element of

$$\left\{i : f_t^i = \text{majority}\left(\{f_t^j : j \in C_t\}\right)\right\}.$$

Theorem

This strategy has regret no more than $\log_2 m$.

Prediction with Expert Advice: Halving

Theorem

The halving strategy has regret no more than $\log_2 m$.

Proof.

If it makes a mistake (that is, $\ell_t(a_t) = 1$), then the minority of $\{f_t^j : j \in C_t\}$ is correct, so at least half of the experts are eliminated:

$$|C_{t+1}| \leq \frac{|C_t|}{2}.$$

And otherwise $|C_{t+1}| \leq |C_t|$ (because $|C_t|$ never increases). Thus,

$$\begin{aligned}\hat{L}_n &= \sum_{t=1}^{n} \ell_t(a_t) \\ &\leq \log_2 \frac{|C_1|}{|C_{n+1}|} = \log_2 m - \log_2 |C_{n+1}| \leq \log_2 m.\end{aligned}$$

Prediction with Expert Advice

The proof follows a pattern we shall see again:
find some measure of progress (here, $|C_t|$) that

- changes monotonically when excess loss is incurred (here, it halves),
- is somehow constrained (here, it cannot fall below 1, because there is an expert who predicts perfectly).

What if there is no perfect expert?

Finite Comparison Class

1. "Prediction with expert advice."
2. With perfect predictions: $\log m$ regret.
3. Exponential weights strategy: $\sqrt{n \log m}$ regret.
4. Refinements and extensions:
   - Exponential weights and $L^* = 0$
   - n unknown
   - L^* unknown
   - *Convex* (versus linear) losses
   - Bayesian interpretation
5. Probabilistic prediction with a finite class.

Prediction with Expert Advice: Mixed Strategies

- We have m experts.
- Allow a mixed strategy, that is, a_t chosen from the simplex Δ^m—the set of distributions on $\{1, \ldots, m\}$,

$$\Delta^m = \left\{ a \in [0,1]^m : \sum_{i=1}^{m} a^i = 1 \right\}.$$

- We can think of the strategy as choosing an element of $\{1, \ldots, m\}$ randomly, according to a distribution a_t. Or we can think of it as playing an element a_t of Δ^m, and incurring the expected loss,

$$\ell_t(a_t) = \sum_{i=1}^{m} a_t^i \ell_t(e_i),$$

where $\ell_t(e_i) \in [0,1]$ is the *loss* incurred by expert i. (e_i denotes the vector with a single 1 in the ith coordinate, and the rest zeros.)

Prediction with Expert Advice: Exponential Weights

- Maintain a set of (unnormalized) weights over experts:

$$w_1^i = 1,$$
$$w_{t+1}^i = w_t^i \exp\left(-\eta \ell_t(e_i)\right).$$

- Here, $\eta > 0$ is a parameter of the algorithm.
- Choose a_t as the normalized vector,

$$a_t = \frac{1}{\sum_{i=1}^m w_t^i} w_t.$$

Prediction with Expert Advice: Exponential Weights

Theorem

The exponential weights strategy with parameter

$$\eta = \sqrt{\frac{8 \ln m}{n}}$$

has regret satisfying

$$\hat{L}_n - L_n^* \leq \sqrt{\frac{n \ln m}{2}}.$$

Exponential Weights: Proof Idea

We use a measure of progress:

$$W_t = \sum_{i=1}^{m} w_t^i.$$

1. W_n grows at least as

$$\exp\left(-\eta \min_i \sum_{t=1}^{n} \ell_t(e_i)\right).$$

2. W_n grows no faster than

$$\exp\left(-\eta \sum_{t=1}^{n} \ell_t(a_t)\right).$$

Exponential Weights: Proof 1

$$
\begin{aligned}
\ln \frac{W_{n+1}}{W_1} &= \ln \left(\sum_{i=1}^{m} w_{n+1}^i \right) - \ln m \\
&= \ln \left(\sum_{i=1}^{m} \exp \left(-\eta \sum_t \ell_t(e_i) \right) \right) - \ln m \\
&\geq \ln \left(\max_i \exp \left(-\eta \sum_t \ell_t(e_i) \right) \right) - \ln m \\
&= -\eta \min_i \left(\sum_t \ell_t(e_i) \right) - \ln m \\
&= -\eta L_n^* - \ln m.
\end{aligned}
$$

Exponential Weights: Proof 2

$$\begin{aligned}
\ln \frac{W_{t+1}}{W_t} &= \ln \left(\frac{\sum_{i=1}^m \exp(-\eta \ell_t(e_i)) w_t^i}{\sum_i w_t^i} \right) \\
&\leq -\eta \frac{\sum_i \ell_t(e_i) w_t^i}{\sum_i w_t^i} + \frac{\eta^2}{8} \\
&= -\eta \ell_t(a_t) + \frac{\eta^2}{8},
\end{aligned}$$

where we have used Hoeffding's inequality:
for a random variable $X \in [a, b]$ and $\lambda \in \mathbb{R}$,

$$\ln \left(\mathbf{E} e^{\lambda X} \right) \leq \lambda \mathbf{E} X + \frac{\lambda^2 (b-a)^2}{8}.$$

Aside: Proof of Hoeffding's inequality

Define

$$A(\lambda) = \log\left(\mathbf{E}e^{\lambda X}\right) = \log\left(\int e^{\lambda x}\, dP(x)\right),$$

where $X \sim P$. Then A is the log normalization of the exponential family random variable X_λ with reference measure P and sufficient statistic x. Since P has bounded support, $A(\lambda) < \infty$ for all λ, and we know that

$$A'(\lambda) = \mathbf{E}(X_\lambda),$$
$$A''(\lambda) = \mathrm{Var}(X_\lambda).$$

Since P has support in $[a, b]$, $\mathrm{Var}(X_\lambda) \leq (b-a)^2/4$. Then a Taylor expansion about $\lambda = 0$ (where X_λ has the same distribution as X) gives

$$A(\lambda) \leq \lambda \mathbf{E}X + \frac{\lambda^2}{8}(b-a)^2.$$

Exponential Weights: Proof

$$-\eta L_n^* - \ln m \leq \ln \frac{W_{n+1}}{W_1} \leq -\eta \hat{L}_n + \frac{n\eta^2}{8}.$$

Thus,

$$\hat{L}_n - L_n^* \leq \frac{\ln m}{\eta} + \frac{\eta n}{8}.$$

Choosing the optimal η gives the result:

Theorem

The exponential weights strategy with parameter $\eta = \sqrt{8 \ln m/n}$ *has regret no more than* $\sqrt{\frac{n \ln m}{2}}$.

Key Points

For a finite set of actions (experts):

- If one action is perfect (i.e., has zero loss), the halving algorithm gives per round regret of

$$\frac{\ln m}{n}.$$

- Exponential weights gives per round regret of

$$O\left(\sqrt{\frac{\ln m}{n}}\right).$$

Prediction with Expert Advice: Refinements

1. Does exponential weights strategy give the faster rate if $L^* = 0$?
2. Do we need to know n to set η?

Prediction with Expert Advice: Refinements

1. Does exponential weights strategy give the faster rate if $L^* = 0$?
Replace Hoeffding:

$$\ln \mathbf{E} e^{\lambda X} \leq \lambda \mathbf{E} X + \frac{\lambda^2}{8},$$

with:

$$\ln \mathbf{E} e^{\lambda X} \leq (e^{\lambda} - 1)\mathbf{E} X.$$

(for $X \in [0, 1]$: linear upper bound on $e^{\lambda X}$).

Exponential Weights: Proof 2

$$\ln \frac{W_{t+1}}{W_t} = \ln \left(\frac{\sum_{i=1}^{m} \exp(-\eta \ell_t(e_i)) w_t^i}{\sum_i w_t^i} \right)$$
$$\leq \left(e^{-\eta} - 1\right) \ell_t(a_t).$$

Thus

$$\hat{L}_n \leq \frac{\eta}{1 - e^{-\eta}} L_n^* + \frac{\ln m}{1 - e^{-\eta}}.$$

For example, if $L_n^* = 0$ and η is large, we obtain a regret bound of roughly $\ln m$ again. And η large is like the halving algorithm (it puts equal weight on all experts that have zero loss so far).

Prediction with Expert Advice: Refinements

2. Do we need to know n to set η?

   - We used the optimal setting $\eta = \sqrt{8 \ln m/n}$. But can this regret bound be achieved uniformly across time?
   - Yes; using a time-varying $\eta_t = \sqrt{8 \ln m/t}$ gives the same rate (worse constants).
   - It is also possible to set η as a function of L_t^*, the best cumulative loss so far, to give the improved bound for small losses uniformly across time (worse constants).

Prediction with Expert Advice: Refinements

3. We could work with arbitrary convex losses on Δ^m:
We defined loss as linear in a:

$$\ell_t(a) = \sum_i a^i \ell_t(e^i).$$

We could replace this with any bounded convex function on Δ^m. The only change in the proof is an equality becomes an inequality:

$$-\eta \frac{\sum_i \ell_t(e_i) w_t^i}{\sum_i w_t^i} \leq -\eta \ell_t(a_t).$$

Prediction with Expert Advice: Refinements

But note that the exponential weights strategy only competes with the *corners* of the simplex:

Theorem

For convex functions $\ell_t : \Delta^m \to [0,1]$*, the exponential weights strategy, with* $\eta = \sqrt{8 \ln m/n}$*, satisfies*

$$\sum_{t=1}^{n} \ell_t(a_t) \leq \min_i \sum_{t=1}^{n} \ell_t(e^i) + \sqrt{\frac{n \ln m}{2}}.$$

Bayesian Interpretation

We can interpret the exponential weights strategy as **Bayesian prediction**:

1. Exponential weights is equivalent to a *Bayesian update* with outcome vector y and model

$$p(y|j) = h(y) \exp(-\eta y^j).$$

2. An easy regret bound for Bayesian prediction shows that its regret, wrt this *scaled log loss*:

$$\ell_{\text{Bayes}}(p, y) = -\frac{1}{\eta} \log \mathbf{E}_{J \sim p} \exp(-\eta y^J),$$

   is no more than $(1/\eta) \log m$.
3. This convex loss matches the linear loss at the corners of the simplex, and (from Hoeffding) differs from the linear loss by no more than $\eta/8$.
4. This implies the earlier regret bound for exponential weights.

Bayesian Update

parameter space: Θ

outcome space: $\mathcal{Y}$

Assume joint: $p(\theta, y) = \underbrace{\pi(\theta)}_{\text{prior}} \underbrace{p(y|\theta)}_{\text{likelihood}}$

predictive distribution: $\hat{p}_{t+1}(y) = p(y|y_1, \ldots, y_t)$

$$= \int p(y|\theta) \underbrace{p(\theta|y_1, \ldots, y_t)}_{p_{t+1}(\theta)} \, d\theta$$

update p_t on Θ: $p_1(\theta) = \pi(\theta)$

$$p_{t+1}(\theta) = \frac{p_t(\theta) p(y_t|\theta)}{\int p_t(\theta') p(y_t|\theta') d\theta'}$$

cumulative log loss: $-\sum_{t=1}^{n} \log \hat{p}_t(y_t).$

Bayesian Interpretation

Suppose that the likelihood is

$$p(y|j) = h(y)\exp(-\eta y_j),$$

for $j = 1, \ldots, m$ and $y \in \mathbb{R}^m$. Then the Bayes update is:

$$p_{t+1}(j) = \frac{1}{Z} p_t(j)\exp(-\eta y_t^j),$$

(where Z is normalization).
If y_t^j is the loss of expert j, this is the exponential weights algorithm.

Performance of Bayesian Prediction

For the log loss, Bayesian prediction competes with any θ, provided that the prior probability of performance better than θ is not too small.

Theorem

For any π, any sequence $y_1, \ldots, y_n$, and any $\theta \in \Theta$,

$$\hat{L}_n \leq L_n(\theta) - \ln\left(\pi(\{\theta' : L_n(\theta') \leq L_n(\theta)\})\right).$$

Performance of Bayesian Prediction: Proof

First,

$$\underbrace{\hat{p}_1(y_1)\cdots\hat{p}_n(y_n)}_{\exp(-\hat{L}_n)} = p(y_1)p(y_2|y_1)\cdots p(y_n|y_1,\ldots,y_{n-1})$$

$$= p(y_1,\ldots,y_n)$$

$$= \int_\Theta \underbrace{p(y_1|\theta)\cdots p(y_n|\theta)}_{\exp(-L_n(\theta))}\, d\pi(\theta),$$

hence

$$\exp(-\hat{L}_n) \geq \int_S \exp(-L_n(\theta))d\pi(\theta)$$

$$\geq \exp(-L_n(\theta_0)) \int_S d\pi(\theta),$$

where $S = \{\theta \in \Theta : L_n(\theta) \leq L_n(\theta_0)\}$. Thus

$$\hat{L}_n \leq L_n(\theta_0) - \ln(\pi(S)).$$

Performance of Bayesian Prediction

For the log loss, Bayesian prediction competes with any θ, provided that the prior probability of performance better than θ is not too small.

Theorem

For any π, any sequence $y_1, \ldots, y_n$, and any $\theta \in \Theta$,

$$\hat{L}_n \leq L_n(\theta) - \ln\left(\pi(\{\theta' : L_n(\theta') \leq L_n(\theta)\})\right).$$

So if $\pi(i) = 1/m$, the exponential weights strategy's log loss is within $\log(m)$ of optimal.
But what is the log loss here?

Bayesian Interpretation

For a posterior p_t on $\{1, \ldots, m\}$, the predicted probability is

$$\hat{p}_t(y) = \mathbf{E}_{J \sim p_t} p(y|J) = c(y) \mathbf{E}_{J \sim p_t} \exp(-\eta y^J),$$

So setting the loss as the negative log of the predicted probability is equivalent to defining the loss of a posterior p_t with outcome y_t as $\eta \ell_{\text{Bayes}}$ with

$$\ell_{\text{Bayes}}(p_t, y_t) = -\frac{1}{\eta} \log \left(\mathbf{E}_{J \sim p_t} \exp(-\eta y_t^J) \right)$$

(ignoring additive constants).
Compare to the linear loss,

$$\ell(p_t, y_t) = \mathbf{E}_{J \sim p_t} y_t^J.$$

These are equal at the corners of the simplex:

$$\ell_{\text{Bayes}}(\delta_j, y) = \ell(\delta_j, y) = y^j.$$

Bayesian Interpretation

The theorem shows that, with this loss and any prior,

$$\hat{L}_{\text{Bayes},n} \leq \min_j \left(\ell(e^j, y_t) - \frac{\log \pi(j)}{\eta} \right).$$

But this is not the linear loss:

$$\ell_{\text{Bayes}}(p_t, y_t) = -\frac{1}{\eta} \log \left(\mathbf{E}_{J \sim p_t} \exp(-\eta y_t^J) \right)$$

versus $\quad \ell(p_t, y_t) = \mathbf{E}_{J \sim p_t} y_t^J.$

They coincide at the corners, $p_t = e^j$, and ℓ_{Bayes} is convex. What is the gap in Jensen's inequality?

Bayesian Interpretation

$$\ell_{\text{Bayes}}(p_t, y_t) = -\frac{1}{\eta} \log \left(\mathbf{E}_{J \sim p_t} \exp(-\eta y_t^J) \right)$$
$$\ell(p_t, y_t) = \mathbf{E}_{J \sim p_t} y_t^J.$$

Hoeffding's inequality for $X \in [a, b]$:

$$-A(-\eta) = -\log \left(\mathbf{E} \exp(-\eta X) \right) \geq \eta \mathbf{E} X - \frac{\eta^2}{8}(b-a)^2,$$

implies

$$\hat{L}_n \leq \hat{L}_{\text{Bayes},n} + \frac{\eta n}{8} \leq \min_j \left(\ell(e^j, y_t) - \frac{\pi(j)}{\eta} \right) + \frac{\eta n}{8}$$
$$= \min_j \ell(e^j, y_t) + \frac{\log m}{\eta} + \frac{\eta n}{8},$$

as before.

Bayesian Interpretation

We can interpret the exponential weights strategy as **Bayesian prediction**:

1. Exponential weights is equivalent to a Bayesian update with model
$$p(y|j) = h(y) \exp(-\eta y^j).$$
2. Easy regret bound for Bayesian prediction shows that its regret wrt the scaled log loss
$$\ell_{\text{Bayes}}(p, y) = -\frac{1}{\eta} \log \mathbf{E}_{J\sim p} \exp(-\eta y^J)$$
is no more than $(1/\eta) \log m$.
3. This convex loss matches the linear loss at the corners of the simplex, and (from Hoeffding) differs from the linear loss by no more than $\eta/8$.
4. This implies the earlier regret bound for exponential weights.

Finite Comparison Class

1. "Prediction with expert advice."
2. With perfect predictions: $\log m$ regret.
3. Exponential weights strategy: $\sqrt{n \log m}$ regret.
4. Refinements and extensions:
   - Exponential weights and $L^* = 0$
   - n unknown
   - L^* unknown
   - *Convex* (versus linear) losses
   - Bayesian interpretation
5. Probabilistic prediction with a finite class.

Probabilistic Prediction Setting

Let's consider a probabilistic formulation of a prediction problem.

- There is a sample of size n drawn i.i.d. from an unknown probability distribution P on $\mathcal{X} \times \mathcal{Y}$: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Some method chooses $\hat{f} : \mathcal{X} \to \mathcal{Y}$.
- It suffers regret

$$\mathbf{E}\ell(\hat{f}(X), Y) - \min_{f \in F} \mathbf{E}\ell(f(X), Y).$$

- Here, F is a class of functions from $\mathcal{X}$ to $\mathcal{Y}$.

Probabilistic Setting: Zero Loss

Theorem

If some $f^* \in F$ *has* $\mathbf{E}\ell(f^*(X), Y) = 0$*, then choosing*

$$\hat{f} \in C_n = \left\{ f \in F : \hat{\mathbf{E}}\ell(f(X), Y) = 0 \right\}$$

leads to regret that is

$$O\left(\frac{\log |F|}{n}\right).$$

Probabilistic Setting: Zero Loss

Proof.

$$\begin{aligned}\Pr(\mathbf{E}\ell(\hat{f}) \geq \epsilon) &\leq \Pr(\exists f \in F : \hat{\mathbf{E}}\ell(f) = 0, \mathbf{E}\ell(f) \geq \epsilon) \\ &\leq |F|(1-\epsilon)^n \\ &\leq |F|e^{-n\epsilon}.\end{aligned}$$

Integrating the tail bound $\Pr(\mathbf{E}\ell(\hat{f})n \geq \ln|F| + x) \leq e^{-x}$ gives $\mathbf{E}\ell(\hat{f}) \leq c\ln|F|/n$. □

Probabilistic Setting

Theorem

Choosing $\hat{f}$ to minimize the empirical risk, $\hat{\mathbf{E}}\ell(f(X), Y)$, leads to regret that is

$$O\left(\sqrt{\frac{\log |F|}{n}}\right).$$

Probabilistic Setting

Proof.

By the triangle inequality and the definition of $\hat{f}$,

$$\mathbf{E}\ell_{\hat{f}} - \min_{f\in F} \mathbf{E}\ell_f \leq 2\mathbf{E}\sup_{f\in F} \left|\hat{\mathbf{E}}\ell_f - \mathbf{E}\ell_f\right|.$$

$$\begin{aligned}
&\mathbf{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{t=1}^{n}(\ell(Y_t, f(X_t)) - P\ell(Y, f(X)))\right| \\
&= \mathbf{E}\sup_{f\in F}\left|\mathbb{P}'\frac{1}{n}\sum_{t=1}^{n}\left(\ell(Y_t, f(X_t)) - \ell(Y_t', f(X_t'))\right)\right| \\
&\leq \mathbf{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell(Y_t, f(X_t)) - \ell(Y_t', f(X_t'))\right)\right| \\
&\leq 2\mathbf{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(Y_t, f(X_t))\right|,
\end{aligned}$$

where (X_t', Y_t') are independent, with same distribution as (X, Y), and ϵ_t are independent Rademacher (uniform ± 1) random variables.

Aside: Rademacher Averages of a Finite Class

Theorem: For $V \subseteq \mathbb{R}^n$, $\mathbf{E} \max_{v \in V} \sum_{i=1}^{n} \epsilon_i v_i \leq \sqrt{2 \ln |V|} \max_{v \in V} \|v\|$.

Proof idea: Hoeffding's inequality.

$$\begin{aligned}
\exp\left(\lambda \mathbf{E} \max_v \sum_i \epsilon_i v_i\right) &\leq \mathbf{E} \exp\left(\lambda \max_v \sum_i \epsilon_i v_i\right) \\
&\leq \sum_v \mathbf{E} \exp\left(\lambda \sum_i \epsilon_i v_i\right) \\
&= \sum_v \prod_i \mathbf{E} \exp(\lambda \epsilon_i v_i) \\
&\leq \sum_v \prod_i \exp(\lambda^2 v_i^2/2) \\
&\leq |V| \exp\left(\lambda^2 \max_v \|v\|^2/2\right).
\end{aligned}$$

Probabilistic Setting

$$\begin{aligned}
\mathbf{E}\ell_{\hat{f}} - \min_{f\in F} \mathbf{E}\ell_f &\le 4\mathbf{E}\sup_{f\in F}\left|\frac{1}{n}\sum_t \epsilon_t \ell_f(X_t, Y_t)\right| \\
&\le 4\max_{X_i,Y_i,f}\sqrt{\sum_t \ell_f(X_i, Y_i)^2}\,\frac{\sqrt{2\log|F|}}{n} \\
&\le 4\sqrt{\frac{2\log|F|}{n}}.
\end{aligned}$$

Probabilistic Setting: Key Points

For a finite function class

- If one function has zero loss, choosing $\hat{f}$ to minimize the empirical risk, $\hat{\mathbf{E}}\ell(f(X), Y)$, gives per round regret of

$$\frac{\ln |F|}{n}.$$

- In any case, $\hat{f}$ has per round regret of

$$O\left(\sqrt{\frac{\ln |F|}{n}}\right).$$

The same as the adversarial setting.

Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
  1. "Prediction with expert advice."
  2. With perfect predictions: $\log m$ regret.
  3. Exponential weights strategy: $\sqrt{n \log m}$ regret.
  4. Refinements and extensions.
  5. Probabilistic prediction with a finite class.
- Online, adversarial versus batch, probabilistic.
- Optimal regret.
- Online convex optimization.

Online to Batch Conversion

- Suppose we have an online strategy that, given observations $\ell_1, \ldots, \ell_{t-1}$, produces $a_t = A(\ell_1, \ldots, \ell_{t-1})$.
- Can we convert this to a method that is suitable for a probabilistic setting? That is, if the ℓ_t are chosen i.i.d., can we use A's choices a_t to come up with an $\hat{a} \in \mathcal{A}$ so that

$$\mathbf{E}\ell_1(\hat{a}) - \min_{a \in \mathcal{A}} \mathbf{E}\ell_1(a)$$

  is small?
- Consider the following simple randomized method:
  1. Pick T uniformly from $\{0, \ldots, n\}$.
  2. Let $\hat{a} = A(\ell_{T+1}, \ldots, \ell_n)$.

Online to Batch Conversion

Theorem

If A has a regret bound of C_{n+1} for sequences of length $n+1$, then for any stationary process generating the $\ell_1, \ldots, \ell_{n+1}$, this method satisfies

$$\mathbf{E}\ell_{n+1}(\hat{a}) - \min_{a \in \mathcal{A}} \mathbf{E}\ell_n(a) \leq \frac{C_{n+1}}{n+1}.$$

(Notice that the expectation averages also over the randomness of the method.)

Online to Batch Conversion

Proof.

$$
\begin{aligned}
\mathbf{E}\ell_{n+1}(\hat{a}) &= \mathbf{E}\ell_{n+1}(A(\ell_{T+1},\ldots,\ell_n)) \\
&= \mathbf{E}\frac{1}{n+1}\sum_{t=0}^{n}\ell_{n+1}(A(\ell_{t+1},\ldots,\ell_n)) \\
&= \mathbf{E}\frac{1}{n+1}\sum_{t=0}^{n}\ell_{n-t+1}(A(\ell_1,\ldots,\ell_{n-t})) \\
&= \mathbf{E}\frac{1}{n+1}\sum_{t=1}^{n+1}\ell_t(A(\ell_1,\ldots,\ell_{t-1})) \\
&\leq \mathbf{E}\frac{1}{n+1}\left(\min_a\sum_{t=1}^{n+1}\ell_t(a)+C_{n+1}\right) \\
&\leq \min_a \mathbf{E}\ell_t(a)+\frac{C_{n+1}}{n+1}.
\end{aligned}
$$

Online to Batch Conversion

- The theorem is for the expectation over the randomness of the method.
- For a high probability result, we could
  1. Choose $\hat{a} = \frac{1}{n}\sum_{t=1}^{n} a_t$, provided $\mathcal{A}$ is convex and the ℓ_t are all convex.
  2. Choose

$$\hat{a} = \arg\min_{a_t} \left(\frac{1}{n-t} \sum_{s=t+1}^{n} \ell_s(a_t) + c\sqrt{\frac{\log(n/\delta)}{n-t}} \right).$$

In both cases, the analysis involves concentration of martingales.

Online to Batch Conversion

Key Point:

- An online strategy with regret bound C_n can be converted to a batch method.
  The regret per trial in the probabilistic setting is bounded by the regret per trial in the adversarial setting.

Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
- Online, adversarial versus batch, probabilistic.
- Optimal regret.
  1. Dual game.
  2. Rademacher averages and sequential Rademacher averages.
  3. Linear games.
- Online convex optimization.

Optimal Regret

Joint work with Jake Abernethy, Alekh Agarwal, Sasha Rakhlin, Karthik Sridharan and Ambuj Tewari.

We have:

- a set of actions $\mathcal{A}$,
- a set of loss functions $\mathcal{L}$.

At time t,

- Player chooses an action a_t from $\mathcal{A}$.
- Adversary chooses $\ell_t : \mathcal{A} \to \mathbb{R}$ from $\mathcal{L}$.
- Player incurs loss $\ell_t(a_t)$.

Regret is the value of the game:

$$V_n(\mathcal{A}, \mathcal{L}) = \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right).$$

Optimal Regret: Dual Game

Theorem

If $\mathcal{A}$ is compact and all ℓ_t are convex, continuous functions, then

$$V_n(\mathcal{A}, \mathcal{L}) = \sup_P \mathbf{E}\left(\sum_{t=1}^{n} \inf_{a_t \in \mathcal{A}} \mathbf{E}\left[\ell_t(a_t) | \ell_1, \ldots, \ell_{t-1}\right] - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a)\right),$$

where the supremum is over joint distributions P over sequences $\ell_1, \ldots, \ell_n$ in $\mathcal{L}^n$.

- As we'll see, this follows from a minimax theorem.
- Dual game: adversary plays first by choosing P.
- Value of the game is the difference between minimal conditional expected loss and minimal empirical loss.
- If P were i.i.d., this would be the difference between the minimal expected loss and the minimal empirical loss.

Optimal Regret: Extensions

- We could replace $\mathcal{L}^n$ by a set of sequences of loss functions:

  $$\ell_1 \in \mathcal{L}_1,\ \ell_2 \in \mathcal{L}_2(\ell_1),\ \ell_3 \in \mathcal{L}_3(\ell_1, \ell_2), \ldots,\ \ell_n \in \mathcal{L}_n(\ell_1, \ell_2, \ldots, \ell_{n-1}).$$

  That is, the constraints on the adversary's choice ℓ_t could depend on previous choices $\ell_1, \ldots, \ell_{t-1}$.
- We can ensure convexity of the ℓ_t by allowing mixed strategies: replace $\mathcal{A}$ by the set of probability distributions P on $\mathcal{A}$ and replace $\ell(a)$ by $\mathbf{E}_{a\sim P}\ell(a)$.

Dual Game: Proof Idea

Theorem (Sion, 1957)

If $\mathcal{A}$ is compact and for every $b \in \mathcal{B}$, $f(\cdot, b)$ is a convex-like,[1] *lower semi-continuous function, and for every $a \in \mathcal{A}$, $f(a, \cdot)$ is concave-like, then*

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} f(a, b) = \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} f(a, b).$$

We'll define $\mathcal{B}$ as the set of probability distributions on $\mathcal{L}$ and $f(a, b) = c + \mathbf{E}[\ell(a) + \phi(\ell)]$, and we'll assume that $\mathcal{A}$ is compact and each $\ell \in \mathcal{L}$ is convex and continuous.

---

[1] *Convex-like* [Fan, 1953]:

$$\forall a_1, a_2 \in \mathcal{A},\ \alpha \in [0, 1],\ \exists a \in \mathcal{A},\ \alpha\ell(a_1) + (1 - \alpha)\ell(a_2) \leq \ell(a).$$

Dual Game: Proof Idea

$$V_n(\mathcal{A},\mathcal{L}) = \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{\ell_n}\left(\sum_{t=1}^{n}\ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)$$
$$= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{P_n}\mathbf{E}\left(\sum_{t=1}^{n}\ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right),$$

because allowing mixed strategies does not help the adversary.

Dual Game: Proof Idea

$$
\begin{aligned}
V_n(\mathcal{A}, \mathcal{L}) &= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right) \\
&= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{P_n} \mathbf{E} \left(\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right) \\
&= \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_{n-1}} \sup_{\ell_{n-1}} \sup_{P_n} \inf_{a_n} \mathbf{E} \left(\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right),
\end{aligned}
$$

by Sion's generalization of von Neumann's minimax theorem.

Dual Game: Proof Idea

$$
\begin{aligned}
V_n(\mathcal{A},\mathcal{L}) &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{\ell_n}\left(\sum_{t=1}^{n}\ell_t(a_t)-\inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\\
&= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{P_n}\mathbf{E}\left(\sum_{t=1}^{n}\ell_t(a_t)-\inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\\
&= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-1}}\sup_{\ell_{n-1}}\sup_{P_n}\inf_{a_n}\mathbf{E}\left(\sum_{t=1}^{n}\ell_t(a_t)-\inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\\
&= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-1}}\sup_{P_{n-1}}\mathbf{E}\left(\sum_{t=1}^{n-1}\ell_t(a_t)+\right.\\
&\qquad\left.\sup_{P_n}\left(\inf_{a_n}\mathbf{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right]-\inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\right),
\end{aligned}
$$

splitting the sum and allowing the adversary a mixed strategy at round $n-1$.

Dual Game: Proof Idea

$$V_n(\mathcal{A},\mathcal{L}) = \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-1}}\sup_{P_{n-1}}\mathbf{E}\left(\sum_{t=1}^{n-1}\ell_t(a_t) + \sup_{P_n}\left(\inf_{a_n}\mathbf{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\right)$$

$$= \inf_{a_1}\sup_{\ell_1}\cdots\sup_{P_{n-1}}\inf_{a_{n-1}}\mathbf{E}\left(\sum_{t=1}^{n-1}\ell_t(a_t) + \sup_{P_n}\left(\inf_{a_n}\mathbf{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\right),$$

applying Sion's minimax theorem again.

Dual Game: Proof Idea

$$V_n(\mathcal{A},\mathcal{L}) = \inf_{a_1}\sup_{\ell_1}\cdots\sup_{P_{n-1}}\inf_{a_{n-1}}\mathbf{E}\left(\sum_{t=1}^{n-1}\ell_t(a_t) + \sup_{P_n}\left(\inf_{a_n}\mathbf{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\right)$$

$$= \inf_{a_1}\sup_{\ell_1}\cdots\sup_{P_{n-2}}\inf_{a_{n-2}}\left(\mathbf{E}\sum_{t=1}^{n-2}\ell_t(a_t) + \sup_{P_{n-1}^n}\mathbf{E}\left(\sum_{t=n-1}^{n}\inf_{a_t}\mathbf{E}\left[\ell_t(a_t)|\ell_1,\ldots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right)\right)$$

$$\vdots$$

$$= \sup_{P}\mathbf{E}\left(\sum_{t=1}^{n}\inf_{a_t}\mathbf{E}\left[\ell_t(a_t)|\ell_1,\ldots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right).$$

Optimal Regret

- Dual game.
- Rademacher averages and sequential Rademacher averages.
- Linear games.

Prediction in Probabilistic Settings

- i.i.d. $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ from $\mathcal{X} \times \mathcal{Y}$.
- Use data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose $f_n : \mathcal{X} \to \mathcal{A}$ with small risk,

$$R(f_n) = P\ell(Y, f_n(X)),$$

ideally not much larger than the minimum risk over some comparison class F:

$$\text{excess risk} = R(f_n) - \inf_{f \in F} R(f).$$

Tools for the analysis of probabilistic problems

For $f_n = \arg\min_{f \in F} \sum_{t=1}^{n} \ell(Y_t, f(X_t))$,

$$R(f_n) - \inf_{f \in F} P\ell(Y, f(X)) \leq 2 \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \ell(Y_t, f(X_t)) - P\ell(Y, f(X)) \right|.$$

So supremum of empirical process, indexed by F, gives upper bound on excess risk.

Tools for the analysis of probabilistic problems

Typically, this supremum is concentrated about

$$\begin{aligned}
&\mathbb{P} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} (\ell(Y_t, f(X_t)) - P\ell(Y, f(X))) \right| \\
&= \mathbb{P} \sup_{f \in F} \left| \mathbb{P}' \frac{1}{n} \sum_{t=1}^{n} \left(\ell(Y_t, f(X_t)) - \ell(Y_t', f(X_t')) \right) \right| \\
&\leq \mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \left(\ell(Y_t, f(X_t)) - \ell(Y_t', f(X_t')) \right) \right| \\
&\leq 2\mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(Y_t, f(X_t)) \right|,
\end{aligned}$$

where (X_t', Y_t') are independent, with same distribution as (X, Y), and ϵ_t are independent Rademacher (uniform ± 1) random variables.

Tools for the analysis of probabilistic problems

That is, for $f_n = \arg\min_{f \in F} \sum_{t=1}^{n} \ell(Y_t, f(X_t))$, with high probability,

$$R(f_n) - \inf_{f \in F} P\ell(Y, f(X)) \leq c\mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(Y_t, f(X_t)) \right|,$$

where ϵ_t are independent Rademacher (uniform ± 1) random variables.

- Rademacher averages capture complexity of $\{(x, y) \mapsto \ell(y, f(x)) : f \in F\}$: they measure how well functions align with a random $(\epsilon_1, \ldots, \epsilon_n)$.
- Rademacher averages are a key tool in analysis of many statistical methods: related to covering numbers (Dudley) and combinatorial dimensions (Vapnik-Chervonenkis, Pollard), for example.
- A related result applies in the online setting...

Optimal Regret and Sequential Rademacher Averages

Theorem

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t \ell_t(a),$$

where $\epsilon_1, \ldots, \epsilon_n$ *are independent Rademacher (uniform* ± 1*-valued) random variables.*

- Compare to the bound involving Rademacher averages in the probabilistic setting:

$$\text{excess risk} \leq c\mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(Y_t, f(X_t)) \right|.$$

- In the adversarial case, the choice of ℓ_t is deterministic, and can depend on $\epsilon_1, \ldots, \epsilon_{t-1}$.

Sequential Rademacher Averages: Proof Idea

$$V_n(\mathcal{A}, \mathcal{L}) = \sup_P \mathbf{E}\left(\sum_{t=1}^{n} \inf_{a_t \in \mathcal{A}} \mathbf{E}\left[\ell_t(a_t)|\ell_1, \ldots, \ell_{t-1}\right] - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a)\right)$$
$$\leq \sup_P \mathbf{E}\left(\sum_{t=1}^{n} \mathbf{E}\left[\ell_t(\hat{a})|\ell_1, \ldots, \ell_{t-1}\right] - \sum_{t=1}^{n} \ell_t(\hat{a})\right),$$

where $\hat{a}$ minimizes $\sum_t \ell_t(a)$.

Sequential Rademacher Averages: Proof Idea

$$
\begin{aligned}
V_n(\mathcal{A}, \mathcal{L}) &= \sup_P \mathbf{E}\left(\sum_{t=1}^{n} \inf_{a_t \in \mathcal{A}} \mathbf{E}\left[\ell_t(a_t)|\ell_1, \ldots, \ell_{t-1}\right] - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a)\right) \\
&\leq \sup_P \mathbf{E}\left(\sum_{t=1}^{n} \mathbf{E}\left[\ell_t(\hat{a})|\ell_1, \ldots, \ell_{t-1}\right] - \sum_{t=1}^{n} \ell_t(\hat{a})\right) \\
&\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\mathbf{E}\left[\ell_t(a)|\ell_1, \ldots, \ell_{t-1}\right] - \ell_t(a)\right).
\end{aligned}
$$

Sequential Rademacher Averages: Proof Idea

$$V_n(\mathcal{A}, \mathcal{L}) \leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\mathbf{E}\left[\ell_t(a) | \ell_1, \ldots, \ell_{t-1}\right] - \ell_t(a) \right)$$

$$= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\mathbf{E}\left[\ell'_t(a) | \ell_1, \ldots, \ell_n\right] - \ell_t(a) \right),$$

where ℓ'_t is a *tangent sequence*: conditionally independent of ℓ_t given $\ell_1, \ldots, \ell_{t-1}$, with the same conditional distribution.

Sequential Rademacher Averages: Proof Idea

$$
\begin{aligned}
V_n(\mathcal{A}, \mathcal{L}) &\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\mathbf{E}\left[\ell_t(a) | \ell_1, \ldots, \ell_{t-1} \right] - \ell_t(a) \right) \\
&= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\mathbf{E}\left[\ell'_t(a) | \ell_1, \ldots, \ell_n \right] - \ell_t(a) \right) \\
&\leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\ell'_t(a) - \ell_t(a) \right),
\end{aligned}
$$

moving the supremum inside the expectation.

Sequential Rademacher Averages: Proof Idea

$$V_n(\mathcal{A}, \mathcal{L}) \leq \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \left(\ell'_t(a) - \ell_t(a) \right)$$

$$= \sup_P \mathbf{E} \sup_{a \in \mathcal{A}} \left(\sum_{t=1}^{n-1} \left(\ell'_t(a) - \ell_t(a) \right) + \epsilon_n \left(\ell'_n(a) - \ell_n(a) \right) \right),$$

for $\epsilon_n \in \{-1, 1\}$, since ℓ'_n has the same conditional distribution, given $\ell_1, \ldots, \ell_{n-1}$, as ℓ_n.

Sequential Rademacher Averages: Proof Idea

$$
\begin{aligned}
V_n(\mathcal{A},\mathcal{L}) &\leq \sup_P \mathbf{E} \sup_{a\in\mathcal{A}} \sum_{t=1}^{n} \left(\ell'_t(a) - \ell_t(a)\right) \\
&= \sup_P \mathbf{E} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n-1} \left(\ell'_t(a) - \ell_t(a)\right) + \epsilon_n \left(\ell'_n(a) - \ell_n(a)\right) \right) \\
&= \sup_P \mathbf{E}_{\ell_1,\ldots,\ell_{n-1}} \mathbf{E}_{\ell_n,\ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n-1} \left(\ell'_t(a) - \ell_t(a)\right) + \right. \\
&\qquad \left. \epsilon_n \left(\ell'_n(a) - \ell_n(a)\right)\right) \\
&\leq \sup_P \mathbf{E}_{\ell_1,\ldots,\ell_{n-1}} \sup_{\ell_n,\ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n-1} \left(\ell'_t(a) - \ell_t(a)\right) + \right. \\
&\qquad \left. \epsilon_n \left(\ell'_n(a) - \ell_n(a)\right)\right).
\end{aligned}
$$

Sequential Rademacher Averages: Proof Idea

$$
\begin{aligned}
V_n(\mathcal{A},\mathcal{L}) &\leq \sup_P \mathbf{E}_{\ell_1,\ldots,\ell_{n-1}} \mathbf{E}_{\ell_n,\ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n-1} \left(\ell'_t(a) - \ell_t(a)\right) + \right. \\
&\qquad \left. \epsilon_n \left(\ell'_n(a) - \ell_n(a)\right)\right) \\
&\leq \sup_P \mathbf{E}_{\ell_1,\ldots,\ell_{n-1}} \sup_{\ell_n,\ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n-1} \left(\ell'_t(a) - \ell_t(a)\right) + \right. \\
&\qquad \left. \epsilon_n \left(\ell'_n(a) - \ell_n(a)\right)\right) \\
&\vdots \\
&\leq \sup_{\ell_1,\ell'_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n,\ell'_n} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n} \epsilon_t \left(\ell'_t(a) - \ell_t(a)\right) \right).
\end{aligned}
$$

Sequential Rademacher Averages: Proof Idea

$$V_n(\mathcal{A},\mathcal{L}) \le \sup_{\ell_1,\ell_1'} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n,\ell_n'} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n} \epsilon_t \left(\ell_t'(a) - \ell_t(a)\right) \right)$$

$$= 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a\in\mathcal{A}} \left(\sum_{t=1}^{n} \epsilon_t \ell_t(a) \right),$$

since the two sums are identical (ϵ_t and $-\epsilon_t$ have the same distribution).

Optimal Regret and Sequential Rademacher Averages

Theorem

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t \ell_t(a),$$

where $\epsilon_1, \ldots, \epsilon_n$ *are independent Rademacher (uniform* ± 1*-valued) random variables.*

- Compare to bound involving Rademacher averages in the probabilistic setting:

$$\text{excess risk} \leq c\mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(Y_t, f(X_t)) \right|.$$

- In the adversarial case, the choice of ℓ_t is deterministic, and can depend on $\epsilon_1, \ldots, \epsilon_{t-1}$: sequential Rademacher averages.

Sequential Rademacher Averages: Example

Consider step functions on $\mathbb{R}$:

$$\begin{aligned} f_a : x &\mapsto 1[x \geq a] \\ \ell_a(y, x) &= 1[f_a(x) \neq y] \\ \mathcal{L} &= \{a \mapsto 1[f_a(x) \neq y] : x \in \mathbb{R},\ y \in \{0, 1\}\}. \end{aligned}$$

Fix a distribution on $\mathbb{R} \times \{\pm 1\}$, and consider the Rademacher averages,

$$\mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_a(Y_t, X_t).$$

Rademacher Averages: Example

For step functions on $\mathbb{R}$, Rademacher averages are:

$$\begin{aligned}
&\mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_a(Y_t, X_t) \\
&= \mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_a(1, X_t) \\
&\leq \sup_{x_t} \mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t 1[x_t < a] \\
&= \mathbf{E} \max_{0 \leq i \leq n+1} \sum_{t=1}^{i} \epsilon_t \\
&= O(\sqrt{n}).
\end{aligned}$$

Sequential Rademacher Averages: Example

Consider the sequential Rademacher averages:

$$\sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t \ell_t(a)$$

$$= \sup_{x_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{x_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t 1[x_t < a].$$

- If $\epsilon_t = 1$, we'd like to choose a such that $x_t < a$.
- If $\epsilon_t = -1$, we'd like to choose a such that $x_t \geq a$.

Sequential Rademacher Averages: Example

Sequential Rademacher averages are

$$\sup_{x_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{x_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t 1[x_t < a].$$

We can choose $x_1 = 0$ and, for $t = 1, \ldots, n$,

$$x_t = \sum_{i=1}^{t-1} 2^{-i} \epsilon_i = x_{t-1} + 2^{-(t-1)} \epsilon_{t-1}.$$

Then if we set

$$a = x_n + 2^{-n} \epsilon_n,$$

we have

$$\epsilon_t 1[x_t < a] = \begin{cases} 1 & \text{if } \epsilon_t = 1, \\ 0 & \text{otherwise,} \end{cases}$$

which is maximal.

Sequential Rademacher Averages: Example

So the sequential Rademacher averages are

$$\sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t \ell_t(a) = \mathbf{E} \sum_{t=1}^{n} 1[\epsilon_t = 1] = \frac{n}{2}.$$

Compare with the Rademacher averages:

$$\mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_a(Y_t, X_t) = O(\sqrt{n}).$$

Optimal Regret

- Dual game.
- Rademacher averages and sequential Rademacher averages.
- Linear games.

Optimal Regret: Linear Games

Loss is $\ell(a) = \langle c, a \rangle$.
Examples:

- Online linear optimization: $\mathcal{L}$ is a set of bounded linear functions on the bounded set $\mathcal{A} \subset \mathbb{R}^d$.
- Prediction with expert advice: $\mathcal{A} = \Delta^m$, $\mathcal{L} = [0,1]^m$.

Optimal Regret: Linear Games

Theorem

For the linear loss class $\{\ell(a) = \langle c, a\rangle : c \in \mathcal{C}\}$*, the regret satisfies*

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}} \mathbf{E} \sup_{a \in \mathcal{A}} \langle Z_n, a\rangle ,$$

where $\mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}$ *is the set of martingales with differences in* $\mathcal{C} \cup -\mathcal{C}$*.*

If $\mathcal{C}$ *is symmetric* $(-\mathcal{C} = \mathcal{C})$*, then*

$$V_n(\mathcal{A}, \mathcal{L}) \geq \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}} \mathbf{E} \sup_{a \in \mathcal{A}} \langle Z_n, a\rangle .$$

Linear Games: Proof Idea

The sequential Rademacher averages can be written

$$\begin{aligned}
&\sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t \ell_t(a) \\
&= \sup_{c_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{c_n} \mathbf{E}_{\epsilon_n} \sup_{a} \left\langle \sum_{t=1}^{n} \epsilon_t c_t, a \right\rangle \\
&\leq \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}} \mathbf{E} \sup_{a} \langle Z_n, a \rangle ,
\end{aligned}$$

where $\mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}$ is the set of martingales with differences in $\mathcal{C} \cup -\mathcal{C}$.

Linear Games: Proof Idea

For the lower bound, consider the duality result:

$$V_n(\mathcal{A},\mathcal{L}) = \sup_P \mathbf{E}\left(\sum_{t=1}^{n} \inf_{a_t\in\mathcal{A}} \mathbf{E}\left[\langle c_t, a_t\rangle | c_1,\ldots,c_{t-1}\right] - \inf_{a\in\mathcal{A}} \sum_{t=1}^{n}\langle c_t, a\rangle\right),$$

where the supremum is over joint distributions of sequences $c_1,\ldots,c_n$. If we restrict P to the set $\mathcal{M}_{\mathcal{C}\cup-\mathcal{C}}$ of martingales with differences in $\mathcal{C}=\mathcal{C}\cup-\mathcal{C}$, the first term is zero and we have

$$\begin{aligned} V_n(\mathcal{A},\mathcal{L}) &\geq \sup_{\{Z_t\}\in\mathcal{M}_{\mathcal{C}\cup-\mathcal{C}}} \mathbf{E} - \inf_{a\in\mathcal{A}}\langle Z_n, a\rangle \\ &= \sup_{\{Z_t\}\in\mathcal{M}_{\mathcal{C}\cup-\mathcal{C}}} \mathbf{E}\sup_{a\in\mathcal{A}}\langle Z_n, a\rangle. \end{aligned}$$

Linear Games: Convex Duality

[Sham Kakade, Karthik Sridharan and Ambuj Tewari, 2009]

Theorem

For the linear loss class $\{\ell(a) = \langle c, a\rangle : c \in \mathcal{C}\}$*, the regret satisfies*

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}} \mathbf{E} \sup_{a \in \mathcal{A}} \langle Z_n, a\rangle .$$

The linear criterion brings to mind a dual norm.
Suppose we have a norm $\|\cdot\|$ defined on $\mathcal{C}$. Then we can view $\mathcal{A}$ as a subset of the dual space of linear functions on $\mathcal{C}$, with the dual norm

$$\|a\|_* = \sup \{\langle c, a\rangle : c \in \mathcal{C},\ \|c\| \leq 1\} .$$

Linear Games: Convex Duality

We will measure the size of $\mathcal{L}$ using $\sup_{c \in \mathcal{C}} \|c\|$.
We will measure the size of $\mathcal{A}$ using $\sup_{a \in \mathcal{A}} R(a)$, for a strongly convex R.
We'll call a function $R : \mathcal{A} \to \mathbb{R}$ σ-strongly convex wrt $\|\cdot\|_*$ if for all $a, b \in \mathcal{A}$ and $\alpha \in [0, 1]$,

$$R(\alpha a + (1-\alpha)b) \leq \alpha R(a) + (1-\alpha)R(b) - \frac{\sigma}{2}\alpha(1-\alpha)\|a-b\|_*^2.$$

Linear Games: Convex Duality

The Legendre dual of R is

$$R^*(c) = \sup_{a \in \mathcal{A}} \left(\langle c, a \rangle - R(a) \right).$$

If $\inf_{c \in \mathcal{C}} R(c) = 0$, then $R^*(0) = 0$.
If R is σ-strongly convex, then R^* is differentiable and σ-smooth wrt $\| \cdot \|$, that is, for all $c, d \in \mathcal{C}$,

$$R^*(c + d) \leq R^*(c) + \langle \nabla R^*(c), d \rangle + \frac{1}{2\sigma} \|d\|^2.$$

Linear Games: Convex Duality

Theorem

For the linear loss class $\{\ell(a) = \langle c, a\rangle : c \in \mathcal{C}\}$, *if* $R : \mathcal{A} \to \mathbb{R}$ *is* σ*-strongly convex, satisfies* $\inf_{c\in\mathcal{C}} R(c) = 0$, *and*

$$\sup_{c\in\mathcal{C}} \|c\| = 1, \qquad \sup_{a\in\mathcal{A}} R(a) = A^2,$$

then the regret satisfies

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\{Z_t\}\in\mathcal{M}_{\mathcal{C}\cup-\mathcal{C}}} \mathbf{E} \sup_{a\in\mathcal{A}} \langle Z_n, a\rangle$$
$$\leq 2A\sqrt{\frac{2n}{\sigma}}.$$

Linear Games: Proof Idea

The definition of the Legendre dual of R is

$$R^*(c) = \sup_{a \in \mathcal{A}} \left(\langle c, a \rangle - R(a) \right).$$

From the definition, for any $\lambda > 0$,

$$\begin{aligned} \mathbf{E} \sup_{a \in \mathcal{A}} \langle a, Z_n \rangle &\leq \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{\lambda} \left(R(a) + R^*(\lambda Z_n) \right) \\ &\leq \frac{A^2}{\lambda} + \frac{\mathbf{E} R^*(\lambda Z_n)}{\lambda}. \end{aligned}$$

Linear Games: Proof Idea

Consider the evolution of $R^*(\lambda Z_t)$. Because R^* is σ-smooth,

$$\begin{aligned}
&\mathbf{E}\left[R^*(\lambda Z_t)|Z_1,\ldots,Z_{t-1}\right] \\
&\leq R^*(\lambda Z_{t-1}) + \mathbf{E}\left[\langle \nabla R^*(\lambda Z_{t-1}), Z_t - Z_{t-1}\rangle \,|Z_1,\ldots,Z_{t-1}\right] + \\
&\qquad \frac{1}{2\sigma}\mathbf{E}\left[\lambda^2\|Z_t - Z_{t-1}\|^2|Z_1,\ldots,,Z_{t-1}\right] \\
&\leq R^*(\lambda Z_{t-1}) + \frac{\lambda^2}{2\sigma}.
\end{aligned}$$

Thus, $R^*(\lambda Z_n) \leq n\lambda^2/(2\sigma)$.

Linear Games: Proof Idea

$$\begin{aligned}\mathbf{E}\sup_{a\in\mathcal{A}}\langle a, Z_n\rangle &\leq \frac{A^2}{\lambda} + \frac{\mathbf{E}R^*(\lambda Z_n)}{\lambda} \\ &\leq \frac{A^2}{\lambda} + \frac{n\lambda}{2\sigma} \\ &= 2A\sqrt{n}2\sigma\end{aligned}$$

for $\lambda = \sqrt{2A^2\sigma/n}$.

Linear Games: Convex Duality

Theorem

For the linear loss class $\{\ell(a) = \langle c, a\rangle : c \in \mathcal{C}\}$*, if* $R : \mathcal{A} \to \mathbb{R}$ *is* σ*-strongly convex, satisfies* $\inf_{c\in\mathcal{C}} R(c) = 0$*, and*

$$\sup_{c\in\mathcal{C}} \|c\| = 1, \qquad \sup_{a\in\mathcal{A}} R(a) = A^2,$$

then the regret satisfies

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2A\sqrt{\frac{2n}{\sigma}}.$$

Linear Games: Examples

The p- and q- norms (with $1/p + 1/q = 1$) on $\mathcal{C} = B_p(1)$ and $\mathcal{A} = B_q(A)$:

$$\begin{aligned} \|c\| &= \|c\|_p, \\ \|a\|_* &= \|a\|_q, \\ R(a) &= \frac{1}{2}\|a\|_q^2 \leq A^2, \\ \sigma &= 2(q-1). \end{aligned}$$

Here, $V_n(\mathcal{A}, \mathcal{L}) \leq A\sqrt{(p-1)n}$.

Linear Games: Examples

The ∞- and 1-norms on $\mathcal{C} = [-1, 1]^d$ and $\mathcal{A} = \Delta^d$:

$$\begin{aligned} \|c\| &= \|c\|_\infty, \\ \|a\|_* &= \|a\|_1, \\ R(a) &= \log d + \sum_i a_i \log a_i \leq \log d, \\ \sigma &= 1. \end{aligned}$$

Here, $V_n(\mathcal{A}, \mathcal{L}) \leq \sqrt{2n \log d}$.
(The bounds in these examples are tight within small constant factors.)

Optimal Regret: Lower Bounds

[Sasha Rakhlin, Karthik Sridharan and Ambuj Tewari, 2010]

- For the case of prediction with absolute loss:

$$\ell_t(a_t) = |y_t - a_t(x_t)|,$$

there are (almost) corresponding lower bounds:

$$\frac{c_1 R_n(\mathcal{A})}{\log^{3/2} n} \leq V_n \leq c_2 R_n(\mathcal{A}),$$

where

$$R_n(\mathcal{A}) = \sup_{x_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{x_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t a(x_t).$$

Optimal Regret: Structural Results

$$R_n(\mathcal{A}) = \sup_{x_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{x_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t a(x_t).$$

It is straightforward to verify that the following properties extend to these *sequential* Rademacher averages:

- $\mathcal{A} \subseteq \mathcal{B}$ implies $R_n(\mathcal{A}) \leq R_n(\mathcal{B})$.
- $R_n(\mathrm{co}(\mathcal{A})) = R_n(\mathcal{A})$.
- $R_n(c\mathcal{A}) = |c| R_n(\mathcal{A})$.
- $R_n(\phi(\mathcal{A})) \leq \|\phi\|_{\mathrm{Lip}} R_n(\mathcal{A})$.
- $R_n(\mathcal{A} + b) = R_n(\mathcal{A})$.

Optimal Regret: Key Points

- Dual game: Adversary chooses a joint distribution to maximize the difference between the minimal conditional expected loss and the minimal empirical loss.
- Upper bound in terms of sequential Rademacher averages.
- Linear games: bound on a martingale using a strongly convex function.

Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
- Online, adversarial versus batch, probabilistic.
- Optimal regret.
- Online convex optimization.
  1. Problem formulation
  2. Empirical minimization fails.
  3. Gradient algorithm.
  4. Regularized minimization
  5. Regret bounds

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Online Convex Optimization

- $\mathcal{A} =$ convex subset of $\mathbb{R}^d$.
- $\mathcal{L} =$ set of convex real functions on $\mathcal{A}$.

For example,

- $\ell_t(a) = (x_t \cdot a - y_t)^2$.
- $\ell_t(a) = |x_t \cdot a - y_t|$.
- $\ell_t(a) = -\log\left(\exp(a' T(y_t) - A(a))\right)$, for $A(a)$ the log normalization of this exponential family, with sufficient statistic $T(y)$.

Online Convex Optimization: Example

Choosing a_t to minimize past losses,
$a_t = \arg\min_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \ell_s(a)$, can fail.
('fictitious play,' 'follow the leader')

- Suppose $\mathcal{A} = [-1, 1]$, $\mathcal{L} = \{a \mapsto v \cdot a : |v| \leq 1\}$.
- Consider the following sequence of losses:

$$
\begin{aligned}
a_1 &= 0, & \ell_1(a) &= \frac{1}{2}a, \\
a_2 &= -1, & \ell_2(a) &= -a, \\
a_3 &= 1, & \ell_3(a) &= a, \\
a_4 &= -1, & \ell_4(a) &= -a, \\
a_5 &= 1, & \ell_5(a) &= a, \\
&\vdots & &\vdots
\end{aligned}
$$

- $a^* = 0$ shows $L_n^* \leq 0$, but $\hat{L}_n = n - 1$.

Online Convex Optimization: Example

- Choosing a_t to minimize past losses can fail.
- The strategy must avoid overfitting, just as in probabilistic settings.
- Similar approaches (regularization; Bayesian inference) are applicable in the online setting.
- First approach: gradient steps.
  Stay close to previous decisions, but move in a direction of improvement.

Online Convex Optimization: Gradient Method

$$
\begin{aligned}
a_1 &\in \mathcal{A}, \\
a_{t+1} &= \Pi_{\mathcal{A}} \left(a_t - \eta \nabla \ell_t(a_t) \right),
\end{aligned}
$$

where $\Pi_{\mathcal{A}}$ is the Euclidean projection on $\mathcal{A}$,

$$\Pi_{\mathcal{A}}(x) = \arg\min_{a \in \mathcal{A}} \|x - a\|.$$

Theorem

For $G = \max_t \|\nabla \ell_t(a_t)\|$ and $D = diam(\mathcal{A})$, the gradient strategy with $\eta = D/(G\sqrt{n})$ has regret satisfying

$$\hat{L}_n - L_n^* \leq GD\sqrt{n}.$$

Online Convex Optimization: Gradient Method

Theorem

For $G = \max_t \|\nabla \ell_t(a_t)\|$ *and* $D = diam(\mathcal{A})$*, the gradient strategy with* $\eta = D/(G\sqrt{n})$ *has regret satisfying*

$$\hat{L}_n - L_n^* \leq GD\sqrt{n}.$$

Example

$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq 1\}$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\| \leq 1\}$.
$D = 2$, $G \leq 1$.
Regret is no more than $2\sqrt{n}$.
(And $O(\sqrt{n})$ is optimal.)

Online Convex Optimization: Gradient Method

Theorem

For $G = \max_t \|\nabla \ell_t(a_t)\|$ and $D = diam(\mathcal{A})$, the gradient strategy with $\eta = D/(G\sqrt{n})$ has regret satisfying

$$\hat{L}_n - L_n^* \leq GD\sqrt{n}.$$

Example

$\mathcal{A} = \Delta^m$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\|_\infty \leq 1\}$.
$D = 2$, $G \leq \sqrt{m}$.
Regret is no more than $2\sqrt{mn}$.

Since competing with the whole simplex is equivalent to competing with the vertices (experts) for linear losses, this is worse than exponential weights ($\sqrt{m}$ versus $\log m$).

Online Convex Optimization: Gradient Method

Proof.

$$\text{Define} \qquad \tilde{a}_{t+1} = a_t - \eta \nabla \ell_t(a_t),$$
$$a_{t+1} = \Pi_{\mathcal{A}}(\tilde{a}_{t+1}).$$

Fix $a \in \mathcal{A}$ and consider the measure of progress $\|a_t - a\|$.

$$\begin{aligned} \|a_{t+1} - a\|^2 &\leq \|\tilde{a}_{t+1} - a\|^2 \\ &= \|a_t - a\|^2 + \eta^2 \|\nabla \ell_t(a_t)\|^2 - 2\eta \nabla_t(a_t) \cdot (a_t - a). \end{aligned}$$

By convexity,

$$\begin{aligned} \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) &\leq \sum_{t=1}^{n} \nabla \ell_t(a_t) \cdot (a_t - a) \\ &\leq \frac{\|a_1 - a\|^2 - \|a_{n+1} - a\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla \ell_t(a_t)\|^2 \end{aligned}$$

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Online Convex Optimization: A Regularization Viewpoint

- Suppose ℓ_t is linear: $\ell_t(a) = g_t \cdot a$.
- Suppose $\mathcal{A} = \mathbb{R}^d$.
- Then minimizing the regularized criterion

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + \frac{1}{2}\|a\|^2 \right)$$

corresponds to the gradient step

$$a_{t+1} = a_t - \eta \nabla \ell_t(a_t).$$

Online Convex Optimization: Regularization

Regularized minimization

Consider the family of strategies of the form:

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right).$$

The regularizer $R : \mathbb{R}^d \to \mathbb{R}$ is strictly convex and differentiable.

Online Convex Optimization: Regularization

Regularized minimization

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right).$$

- R keeps the sequence of a_ts stable: it diminishes ℓ_t's influence.
- We can view the choice of a_{t+1} as trading off two competing forces: making $\ell_t(a_{t+1})$ small, and keeping a_{t+1} close to a_t.
- This is a perspective that motivated many algorithms in the literature. We'll investigate why regularized minimization can be viewed this way.

Properties of Regularization Methods

In the unconstrained case ($\mathcal{A} = \mathbb{R}^d$), regularized minimization is equivalent to minimizing the latest loss and the distance to the previous decision. The appropriate notion of distance is the Bregman divergence $D_{\Phi_{t-1}}$:
Define

$$\Phi_0 = R,$$
$$\Phi_t = \Phi_{t-1} + \eta \ell_t,$$

so that

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right)$$
$$= \arg\min_{a \in \mathcal{A}} \Phi_t(a).$$

Bregman Divergence

Definition

For a strictly convex, differentiable $\Phi : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence wrt Φ is defined, for $a, b \in \mathbb{R}^d$, as

$$D_\Phi(a, b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a - b)) .$$

$D_\Phi(a, b)$ is the difference between $\Phi(a)$ and the value at a of the linear approximation of Φ about b.

Bregman Divergence

$$D_\Phi(a, b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a - b)) .$$

Example

For $a \in \mathbb{R}^d$, the squared euclidean norm, $\Phi(a) = \frac{1}{2}\|a\|^2$, has

$$\begin{aligned} D_\Phi(a, b) &= \frac{1}{2}\|a\|^2 - \left(\frac{1}{2}\|b\|^2 + b \cdot (a - b)\right) \\ &= \frac{1}{2}\|a - b\|^2, \end{aligned}$$

the squared euclidean norm.

Bregman Divergence

$$D_\Phi(a, b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a - b)) .$$

Example

For $a \in [0, \infty)^d$, the unnormalized negative entropy, $\Phi(a) = \sum_{i=1}^d a_i (\ln a_i - 1)$, has

$$\begin{aligned} D_\Phi(a, b) &= \sum_i (a_i(\ln a_i - 1) - b_i(\ln b_i - 1) - \ln b_i(a_i - b_i)) \\ &= \sum_i \left(a_i \ln \frac{a_i}{b_i} + b_i - a_i \right), \end{aligned}$$

the unnormalized KL divergence.
Thus, for $a \in \Delta^d$, $\Phi(a) = \sum_i a_i \ln a_i$ has

$$D_\phi(a, b) = \sum_i a_i \ln \frac{a_i}{b_i}.$$

Bregman Divergence

When the domain of Φ is $\mathcal{A} \subset \mathbb{R}^d$, in addition to differentiability and strict convexity, we make two more assumptions:

- The interior of $\mathcal{A}$ is convex,
- For a sequence approaching the boundary of $\mathcal{A}$, $\|\nabla\Phi(a_n)\| \to \infty$.

We say that such a Φ is a *Legendre function*.

Bregman Divergence

Properties:

1. $D_\Phi \geq 0$, $D_\Phi(a, a) = 0$.
2. $D_{A+B} = D_A + D_B$.
3. *Bregman projection*, $\Pi_{\mathcal{A}}^{\Phi}(b) = \arg\min_{a \in \mathcal{A}} D_\Phi(a, b)$ is uniquely defined for closed, convex $\mathcal{A}$.
4. *Generalized Pythagorus*: for closed, convex $\mathcal{A}$, $a^* = \Pi_{\mathcal{A}}^{\Phi}(b)$, and $a \in \mathcal{A}$,

$$D_\Phi(a, b) \geq D_\Phi(a, a^*) + D_\Phi(a^*, b).$$

5. $\nabla_a D_\Phi(a, b) = \nabla\Phi(a) - \nabla\Phi(b)$.
6. For ℓ linear, $D_{\Phi+\ell} = D_\Phi$.
7. For Φ^* the Legendre dual of Φ,

$$\nabla\Phi^* = (\nabla\Phi)^{-1},$$
$$D_\Phi(a, b) = D_{\Phi^*}(\nabla\phi(b), \nabla\phi(a)).$$

Legendre Dual

For a Legendre function $\Phi : \mathcal{A} \to \mathbb{R}$, the Legendre dual is

$$\Phi^*(u) = \sup_{v \in \mathcal{A}} (u \cdot v - \Phi(v)) .$$

- Φ^* is Legendre.
- $\mathrm{dom}(\Phi^*) = \nabla\Phi(\mathrm{int\ dom}\ \Phi)$.
- $\nabla\Phi^* = (\nabla\Phi)^{-1}$.
- $D_\Phi(a, b) = D_{\Phi^*}(\nabla\phi(b), \nabla\phi(a))$.
- $\Phi^{**} = \Phi$.

Legendre Dual

Example

For $\Phi = \frac{1}{2}\|\cdot\|_p^2$, the Legendre dual is $\Phi^* = \frac{1}{2}\|\cdot\|_q^2$, where $1/p + 1/q = 1$.

Example

For $\Phi(a) = \sum_{i=1}^{d} e^{a_i}$,

$$\nabla\Phi(a) = (e^{a_1}, \ldots, e^{a_d})',$$

so

$$(\nabla\Phi)^{-1}(u) = \nabla\Phi^*(u) = (\ln u_1, \ldots, \ln u_d)',$$

and $\Phi^*(u) = \sum_i u_i(\ln u_i - 1)$.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Properties of Regularization Methods

In the unconstrained case ($\mathcal{A} = \mathbb{R}^d$), regularized minimization is equivalent to minimizing the latest loss and the distance (Bregman divergence) to the previous decision.

Theorem

Define $\tilde{a}_1$ *via* $\nabla R(\tilde{a}_1) = 0$, *and set*

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\eta \ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) \right).$$

Then

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right).$$

Properties of Regularization Methods

Proof.

By the definition of Φ_t,

$$\eta \ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) = \Phi_t(a) - \Phi_{t-1}(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t).$$

The derivative wrt a is

$$\begin{aligned} &\nabla \Phi_t(a) - \nabla \Phi_{t-1}(a) + \nabla_a D_{\Phi_{t-1}}(a, \tilde{a}_t) \\ &= \nabla \Phi_t(a) - \nabla \Phi_{t-1}(a) + \nabla \Phi_{t-1}(a) - \nabla \Phi_{t-1}(\tilde{a}_t) \end{aligned}$$

Setting to zero shows that

$$\nabla \Phi_t(\tilde{a}_{t+1}) = \nabla \Phi_{t-1}(\tilde{a}_t) = \cdots = \nabla \Phi_0(\tilde{a}_1) = \nabla R(\tilde{a}_1) = 0,$$

So $\tilde{a}_{t+1}$ minimizes Φ_t. □

Properties of Regularization Methods

Constrained minimization is equivalent to unconstrained minimization, followed by Bregman projection:

Theorem

For

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \Phi_t(a),$$
$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a),$$

we have

$$a_{t+1} = \Pi_{\mathcal{A}}^{\Phi_t}(\tilde{a}_{t+1}).$$

Properties of Regularization Methods

Proof.

Let a'_{t+1} denote $\Pi_{\mathcal{A}}^{\Phi_t}(\tilde{a}_{t+1})$. First, by definition of a_{t+1},

$$\Phi_t(a_{t+1}) \leq \Phi_t(a'_{t+1}).$$

Conversely,

$$D_{\Phi_t}(a'_{t+1}, \tilde{a}_{t+1}) \leq D_{\Phi_t}(a_{t+1}, \tilde{a}_{t+1}).$$

But $\nabla\Phi_t(\tilde{a}_{t+1}) = 0$, so

$$D_{\Phi_t}(a, \tilde{a}_{t+1}) = \Phi_t(a) - \Phi_t(\tilde{a}_{t+1}).$$

Thus, $\Phi_t(a'_{t+1}) \leq \Phi_t(a_{t+1})$. □

Properties of Regularization Methods

Example

For linear ℓ_t, regularized minimization is equivalent to minimizing the last loss plus the Bregman divergence wrt R to the previous decision:

$$\begin{aligned}
&\arg\min_{a\in\mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right) \\
&= \Pi_{\mathcal{A}}^{R} \left(\arg\min_{a\in\mathbb{R}^d} \left(\eta \ell_t(a) + D_R(a, \tilde{a}_t) \right) \right),
\end{aligned}$$

because adding a linear function to Φ does not change D_Φ.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Properties of Regularization Methods: Linear Loss

We can replace ℓ_t by $\nabla \ell_t(a_t)$, and this leads to an upper bound on regret.

Theorem

Any strategy for online linear optimization, with regret satisfying

$$\sum_{t=1}^{n} g_t \cdot a_t - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} g_t \cdot a \leq C_n(g_1, \ldots, g_n)$$

can be used to construct a strategy for online convex optimization, with regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \leq C_n(\nabla \ell_1(a_1), \ldots, \nabla \ell_n(a_n)).$$

Proof.

Convexity implies $\ell_t(a_t) - \ell_t(a) \leq \nabla \ell_t(a_t) \cdot (a_t - a)$. □

Properties of Regularization Methods: Linear Loss

Key Point:

We can replace ℓ_t by $\nabla \ell_t(a_t)$, and this leads to an upper bound on regret.

Thus, we can work with linear ℓ_t.

Regularization Methods: Mirror Descent

Regularized minimization for linear losses can be viewed as mirror descent—taking a gradient step in a dual space:

Theorem

The decisions

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\eta \sum_{s=1}^{t} g_s \cdot a + R(a) \right)$$

can be written

$$\tilde{a}_{t+1} = (\nabla R)^{-1} \left(\nabla R(\tilde{a}_t) - \eta g_t \right).$$

This corresponds to first mapping from $\tilde{a}_t$ through ∇R, then taking a step in the direction $-g_t$, then mapping back through $(\nabla R)^{-1} = \nabla R^*$ to $\tilde{a}_{t+1}$.

Regularization Methods: Mirror Descent

Proof.

For the unconstrained minimization, we have

$$\nabla R(\tilde{a}_{t+1}) = -\eta \sum_{s=1}^{t} g_s,$$

$$\nabla R(\tilde{a}_t) = -\eta \sum_{s=1}^{t-1} g_s,$$

so $\nabla R(\tilde{a}_{t+1}) = \nabla R(\tilde{a}_t) - \eta g_t$, which can be written

$$\tilde{a}_{t+1} = \nabla R^{-1} \left(\nabla R(\tilde{a}_t) - \eta g_t \right).$$

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization and Bregman divergences
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Online Convex Optimization: Regularization

Regularized minimization

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right).$$

The regularizer $R : \mathbb{R}^d \to \mathbb{R}$ is strictly convex and differentiable.

Regularization Methods: Regret

Theorem

For $\mathcal{A} = \mathbb{R}^d$, regularized minimization suffers regret against any $a \in \mathcal{A}$ of

$$\sum_{t=1}^{n} \ell_t(a_t) - \sum_{t=1}^{n} \ell_t(a) = \frac{D_R(a, a_1) - D_{\Phi_n}(a, a_{n+1})}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}),$$

and thus

$$\hat{L}_n \leq \inf_{a \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \ell_t(a) + \frac{D_R(a, a_1)}{\eta} \right) + \frac{1}{\eta} \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}).$$

So the sizes of the steps $D_{\Phi_t}(a_t, a_{t+1})$ determine the regret bound.

Regularization Methods: Regret

Theorem

For $\mathcal{A} = \mathbb{R}^d$, regularized minimization suffers regret

$$\hat{L}_n \leq \inf_{a \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \ell_t(a) + \frac{D_R(a, a_1)}{\eta} \right) + \frac{1}{\eta} \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}).$$

Notice that we can write

$$\begin{aligned} D_{\Phi_t}(a_t, a_{t+1}) &= D_{\Phi_t^*}(\nabla \Phi_t(a_{t+1}), \nabla \Phi_t(a_t)) \\ &= D_{\Phi_t^*}(0, \nabla \Phi_{t-1}(a_t) + \eta \nabla \ell_t(a_t)) \\ &= D_{\Phi_t^*}(0, \eta \nabla \ell_t(a_t)). \end{aligned}$$

So it is the size of the gradient steps, $D_{\Phi_t^*}(0, \eta \nabla \ell_t(a_t))$, that determines the regret.

Regularization Methods: Regret Bounds

Example

Suppose $R = \frac{1}{2}\|\cdot\|^2$. Then we have

$$\hat{L}_n \leq L_n^* + \frac{\|a^* - a_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\|g_t\|^2.$$

And if $\|g_t\| \leq G$ and $\|a^* - a_1\| \leq D$, choosing η appropriately gives $\hat{L}_n - L_n^* \leq DG\sqrt{n}$.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization and Bregman divergences
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Regularization Methods: Regret Bounds

Seeing the future gives small regret:

Theorem

For regularized minimization, that is,

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right),$$

for all $a \in \mathcal{A}$,

$$\sum_{t=1}^{n} \ell_t(a_{t+1}) - \sum_{t=1}^{n} \ell_t(a) \leq \frac{1}{\eta}(R(a) - R(a_1)).$$

Regularization Methods: Regret Bounds

Proof.

Since a_{t+1} minimizes Φ_t,

$$\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \geq \eta \sum_{s=1}^{t} \ell_s(a_{t+1}) + R(a_{t+1})$$

$$= \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_{t+1}) + R(a_{t+1})$$

$$\geq \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_t) + R(a_t)$$

$$\vdots$$

$$\geq \eta \sum_{s=1}^{t} \ell_s(a_{s+1}) + R(a_1).$$

Regularization Methods: Regret Bounds

Theorem

For all $a \in \mathcal{A}$,

$$\sum_{t=1}^{n} \ell_t(a_{t+1}) - \sum_{t=1}^{n} \ell_t(a) \leq \frac{1}{\eta}(R(a) - R(a_1)).$$

Thus, if a_t and a_{t+1} are close, then regret is small:

Corollary

For all $a \in \mathcal{A}$,

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a_{t+1})) + \frac{1}{\eta} (R(a) - R(a_1)) .$$

So how can we control the increments $\ell_t(a_t) - \ell_t(a_{t+1})$?

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent and Bregman divergence from previous
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Regularization Methods: Regret Bounds

Definition

We say R is strongly convex wrt a norm $\|\cdot\|$ if, for all a, b,

$$R(a) \geq R(b) + \nabla R(b) \cdot (a - b) + \frac{1}{2}\|a - b\|^2.$$

For linear losses and strongly convex regularizers, the dual norm of the gradient is small:

Theorem

If R is strongly convex wrt a norm $\|\cdot\|$, and $\ell_t(a) = g_t \cdot a$, then

$$\|a_t - a_{t+1}\| \leq \eta \|g_t\|_*,$$

where $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$:

$$\|v\|_* = \sup\{|v \cdot a| : a \in \mathcal{A}, \|a\| \leq 1\}.$$

Regularization Methods: Regret Bounds

Proof.

$$R(a_t) \geq R(a_{t+1}) + \nabla R(a_{t+1}) \cdot (a_t - a_{t+1}) + \frac{1}{2}\|a_t - a_{t+1}\|^2,$$

$$R(a_{t+1}) \geq R(a_t) + \nabla R(a_t) \cdot (a_{t+1} - a_t) + \frac{1}{2}\|a_t - a_{t+1}\|^2.$$

Combining,

$$\|a_t - a_{t+1}\|^2 \leq (\nabla R(a_t) - \nabla R(a_{t+1})) \cdot (a_t - a_{t+1})$$

Hence,

$$\|a_t - a_{t+1}\| \leq \|\nabla R(a_t) - \nabla R(a_{t+1})\|_* = \|\eta g_t\|_*.$$

□

Regularization Methods: Regret Bounds

This leads to the regret bound:

Corollary

For linear losses, if R is strongly convex wrt $\|\cdot\|$*, then for all* $a \in \mathcal{A}$*,*

$$\sum_{t=1}^{n}(\ell_t(a_t) - \ell_t(a)) \leq \eta \sum_{t=1}^{n} \|g_t\|_*^2 + \frac{1}{\eta}\left(R(a) - R(a_1)\right).$$

Thus, for $\|g_t\|_* \leq G$ and $R(a) - R(a_1) \leq D^2$, choosing η appropriately gives regret no more than $2GD\sqrt{n}$.

Regularization Methods: Regret Bounds

Example

Consider $R(a) = \frac{1}{2}\|a\|^2$, $a_1 = 0$, and $\mathcal{A}$ contained in a Euclidean ball of diameter D.
Then R is strongly convex wrt $\|\cdot\|$ and $\|\cdot\|_* = \|\cdot\|$. And the mapping between primal and dual spaces is the identity.
So if $\sup_{a\in\mathcal{A}} \|\nabla \ell_t(a)\| \leq G$, then regret is no more than $2GD\sqrt{n}$.

Regularization Methods: Regret Bounds

Example

Consider $\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$. Then the mapping between primal and dual spaces is $\nabla R(a) = \ln(a)$ (component-wise). And the divergence is the KL divergence,

$$D_R(a, b) = \sum_i a_i \ln(a_i / b_i).$$

And R is strongly convex wrt $\|\cdot\|_1$.
Suppose that $\|g_t\|_\infty \leq 1$. Also, $R(a) - R(a_1) \leq \ln m$, so the regret is no more than $2\sqrt{n \ln m}$.

Regularization Methods: Regret Bounds

Example

$\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$.
What are the updates?

$$\begin{aligned} a_{t+1} &= \Pi^R_{\mathcal{A}}(\tilde{a}_{t+1}) \\ &= \Pi^R_{\mathcal{A}}(\nabla R^*(\nabla R(\tilde{a}_t) - \eta g_t)) \\ &= \Pi^R_{\mathcal{A}}(\nabla R^*(\ln(\tilde{a}_t \exp(-\eta g_t))) \\ &= \Pi^R_{\mathcal{A}}(\tilde{a}_t \exp(-\eta g_t)), \end{aligned}$$

where the ln and exp functions are applied component-wise.
This is exponentiated gradient: mirror descent with $\nabla R = \ln$.
It is easy to check that the projection corresponds to normalization, $\Pi^R_{\mathcal{A}}(\tilde{a}) = \tilde{a}/\|a\|_1$.

Regularization Methods: Regret Bounds

Notice that when the losses are linear, exponentiated gradient is exactly the exponential weights strategy we discussed for a finite comparison class.
Compare $R(a) = \sum_i a_i \ln a_i$ with $R(a) = \frac{1}{2}\|a\|^2$,
for $\|g_t\|_\infty \leq 1$, $\mathcal{A} = \Delta^m$:

$O(\sqrt{n \ln m})$ versus $O(\sqrt{mn})$.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent and Bregman divergence from previous
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Regularization Methods: Extensions

- Instead of

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) \right),$$

  we can use

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \ell_t(a) + D_{\Phi_{t-1}}(a, a_t) \right).$$

  And analogous results apply. For instance, this is the approach used by the first gradient method we considered.
- We can get faster rates with stronger assumptions on the losses...

Regularization Methods: Varying η

Theorem

Define

$$a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \eta_t \ell_t(a) + R(a) \right).$$

For any $a \in \mathbb{R}^d$,

$$\hat{L}_n - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} \left(D_{\Phi_t}(a_t, a_{t+1}) + D_{\Phi_{t-1}}(a, a_t) - D_{\Phi_t}(a, a_{t+1}) \right).$$

If we linearize the ℓ_t, we have

$$\hat{L}_n - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} \left(D_R(a_t, a_{t+1}) + D_R(a, a_t) - D_R(a, a_{t+1}) \right).$$

But what if ℓ_t are strongly convex?

Regularization Methods: Strongly Convex Losses

Theorem

If ℓ_t is σ-strongly convex wrt R, that is, for all $a, b \in \mathbb{R}^d$,

$$\ell_t(a) \geq \ell_t(b) + \nabla \ell_t(b) \cdot (a - b) + \frac{\sigma}{2} D_R(a, b),$$

then for any $a \in \mathbb{R}^d$, this strategy with $\eta_t = \frac{2}{t\sigma}$ has regret

$$\hat{L}_n - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, a_{t+1}).$$

Strongly Convex Losses: Proof idea

$$
\begin{aligned}
&\sum_{t=1}^{n} \left(\ell_t(a_t) - \ell_t(a)\right) \\
&\leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t)\right) \\
&\leq \sum_{t=1}^{n} \frac{1}{\eta_t} \left(D_R(a_t, a_{t+1}) + D_R(a, a_t) - D_R(a, a_{t+1}) - \frac{\eta_t \sigma}{2} D_R(a, a_t)\right) \\
&\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, a_{t+1}) + \sum_{t=2}^{n} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\sigma}{2}\right) D_R(a, a_t) \\
&\qquad + \left(\frac{1}{\eta_1} - \frac{\sigma}{2}\right) D_R(a, a_1).
\end{aligned}
$$

And choosing η_t appropriately eliminates the second and third terms.

Strongly Convex Losses

Example

For $R(a) = \frac{1}{2}\|a\|^2$, we have

$$\hat{L}_n - L_n^* \leq \frac{1}{2}\sum_{t=1}^{n} \frac{1}{\eta_t}\|\eta_t \nabla \ell_t\|^2 \leq \sum_{t=1}^{n} \frac{G^2}{t\sigma} = O\left(\frac{G^2}{\sigma}\log n\right).$$

Strongly Convex Losses

Key Point: When the loss is strongly convex wrt the regularizer, the regret rate can be faster; in the case of quadratic R (and ℓ_t), it is $O(\log n)$, versus $O(\sqrt{n})$.

Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
- Online, adversarial versus batch, probabilistic.
- Optimal regret.
- Online convex optimization.