Topics in Prediction and Learning Lectures 2 and 3: Online Convex Optimization

Peter Bartlett

Computer Science and Statistics
University of California at Berkeley

Mathematical Sciences
Queensland University of Technology

27 February–9 March, 2017
CREST, ENSAE

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses loss $\ell_t \in \mathcal{L}$.

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses loss $\ell_t \in \mathcal{L}$.
3. Player incurs loss $\ell_t(a_t)$.

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses loss $\ell_t \in \mathcal{L}$.
3. Player incurs loss $\ell_t(a_t)$.

Player's aim:

Minimize *regret:*

$$R_n := \sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a).$$

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses loss $\ell_t \in \mathcal{L}$.
3. Player incurs loss $\ell_t(a_t)$.

Player's aim:

Minimize *regret* wrt comparison $\mathcal{C}$:

$$R_n := \sum_{t=1}^{n} \ell_t(a_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell_t(\hat{a}_t).$$

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Quadratic loss: $\ell_t(a) = \|x_t - a\|^2$.

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Quadratic loss: $\ell_t(a) = \|x_t - a\|^2$.
- Linear regression: $\ell_t(a) = (x_t \cdot a - y_t)^2$.

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Quadratic loss: $\ell_t(a) = \|x_t - a\|^2$.
- Linear regression: $\ell_t(a) = (x_t \cdot a - y_t)^2$.
- Absolute loss linear regression: $\ell_t(a) = |x_t \cdot a - y_t|$.

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Quadratic loss: $\ell_t(a) = \|x_t - a\|^2$.
- Linear regression: $\ell_t(a) = (x_t \cdot a - y_t)^2$.
- Absolute loss linear regression: $\ell_t(a) = |x_t \cdot a - y_t|$.
- Prediction with expert advice: $\ell_t(a) = w_t^\top a$ $(\mathcal{A} = \Delta^m)$.

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Shortest path: $\ell_t(a) = w_t^\top a$ ($\mathcal{A} =$ flow, $w_t =$ edge weights).

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Shortest path: $\ell_t(a) = w_t^\top a$ ($\mathcal{A} =$ flow, $w_t =$ edge weights).
- Portfolio optimization: $\ell_t(a) = -\log(r_t^\top a)$ ($\mathcal{A} = \Delta^m$).

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Shortest path: $\ell_t(a) = w_t^\top a$ ($\mathcal{A} =$ flow, $w_t =$ edge weights).
- Portfolio optimization: $\ell_t(a) = -\log(r_t^\top a)$ ($\mathcal{A} = \Delta^m$).
- Collaborative filtering: $\ell_t(A) = (x_t - A_{i_t,j_t})^2$. ($\mathcal{A} = \mathbb{R}^{m\times n}$).

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Shortest path: $\ell_t(a) = w_t^\top a$ ($\mathcal{A} =$ flow, $w_t =$ edge weights).
- Portfolio optimization: $\ell_t(a) = -\log(r_t^\top a)$ ($\mathcal{A} = \Delta^m$).
- Collaborative filtering: $\ell_t(A) = (x_t - A_{i_t,j_t})^2$. ($\mathcal{A} = \mathbb{R}^{m\times n}$).
- SVM: $\ell_t(A) = \left(1 - y_t x_t^\top a\right)_+ + \lambda\|a\|^2$. ($\mathcal{A} =$ RKHS).

Online Prediction as a Zero-Sum Game

Online Convex Optimization

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

Examples

- Shortest path: $\ell_t(a) = w_t^\top a$ ($\mathcal{A} =$ flow, $w_t =$ edge weights).
- Portfolio optimization: $\ell_t(a) = -\log(r_t^\top a)$ ($\mathcal{A} = \Delta^m$).
- Collaborative filtering: $\ell_t(A) = (x_t - A_{i_t,j_t})^2$. ($\mathcal{A} = \mathbb{R}^{m \times n}$).
- SVM: $\ell_t(A) = \left(1 - y_t x_t^\top a\right)_+ + \lambda \|a\|^2$. ($\mathcal{A} =$ RKHS).
- Density estimation: $\ell_t(a) = -\log\left(\exp(a' T(y_t) - A(a))\right)$, for exponential family with sufficient statistic $T(y)$.

Online convex optimization

1. Binary prediction

Online convex optimization

1. Binary prediction

2. General online convex

Online convex optimization

1. Binary prediction

2. General online convex
3. Minimax strategies

Online convex optimization

1. Binary prediction
    - With (perfect) expert advice
2. General online convex
3. Minimax strategies

Online convex optimization

1. Binary prediction
    - With (perfect) expert advice
    - Minimax strategy
2. General online convex
3. Minimax strategies

Online convex optimization

1. Binary prediction
    - With (perfect) expert advice
    - Minimax strategy
    - With imperfect experts: exponential weights
2. General online convex
3. Minimax strategies

Binary Prediction with Expert Advice

- Suppose we are predicting whether it will rain tomorrow.

Binary 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.

Binary 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.
- Can we ensure that we predict almost as well as the best expert?

Binary 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.
- Can we ensure that we predict almost as well as the best expert?
- We'll consider two settings: voting and prediction.

Binary Prediction with Expert Advice

Voting

The player votes for a mixture of experts:

Binary Prediction with Expert Advice

Voting

The player votes for a mixture of experts:
we set $\mathcal{A} = \Delta^m$, the probability simplex on $\{1, \ldots, m\}$, and the loss function at time t is $\ell_t(a) = |a^\top f_t - y_t|$, where $f_t \in \{0, 1\}^m$ are the forecasts of the experts and $y_t \in \{0, 1\}$ is the outcome.

Binary Prediction with Expert Advice

Voting

The player votes for a mixture of experts:
we set $\mathcal{A} = \Delta^m$, the probability simplex on $\{1, \ldots, m\}$, and the loss function at time t is $\ell_t(a) = |a^\top f_t - y_t|$, where $f_t \in \{0, 1\}^m$ are the forecasts of the experts and $y_t \in \{0, 1\}$ is the outcome.

Prediction

The player votes for a mixture of experts, but the vote can depend on their forecasts:

Binary Prediction with Expert Advice

Voting

The player votes for a mixture of experts:
we set $\mathcal{A} = \Delta^m$, the probability simplex on $\{1, \ldots, m\}$, and the loss function at time t is $\ell_t(a) = |a^\top f_t - y_t|$, where $f_t \in \{0,1\}^m$ are the forecasts of the experts and $y_t \in \{0,1\}$ is the outcome.

Prediction

The player votes for a mixture of experts, but the vote can depend on their forecasts: we set $\mathcal{A} = (\Delta^m)^{\{0,1\}^m}$, and the loss function at time t is $\ell_t(a) = |a(f_t)^\top f_t - y_t|$.

Binary Prediction with Expert Advice

Voting

The player votes for a mixture of experts:
we set $\mathcal{A} = \Delta^m$, the probability simplex on $\{1, \ldots, m\}$, and the loss function at time t is $\ell_t(a) = |a^\top f_t - y_t|$, where $f_t \in \{0,1\}^m$ are the forecasts of the experts and $y_t \in \{0,1\}$ is the outcome.

Prediction

The player votes for a mixture of experts, but the vote can depend on their forecasts: we set $\mathcal{A} = (\Delta^m)^{\{0,1\}^m}$, and the loss function at time t is $\ell_t(a) = |a(f_t)^\top f_t - y_t|$.
The comparison class $\mathcal{C}$ is the set of constant functions. (That is, $a \in \mathcal{C}$ has $p \in \Delta^m$ so that for all $f \in \{0,1\}^m$, $a(f) = p$.)

Binary Prediction with Expert Advice

Prediction allows the player to see how the experts' predictions compare before making a prediction.

Binary Prediction with Expert Advice

Prediction allows the player to see how the experts' predictions compare before making a prediction.

We write $\ell_t(e_i) \in \{0, 1\}$ for the loss incurred by expert i, where $e_i \in \Delta^m$ is zero in all but the ith coordinate. and $\ell_t(e_i) \in \{0, 1\}$ is the indicator for expert i making an incorrect forecast at time t.

We can interpret any $a \in \Delta^m$ equivalently as a prediction, $\hat{y}_t = a^\top f_t \in [0, 1]$. And we can view $\hat{y}_t$ either as the expectation of a random $\{0, 1\}$-valued prediction where the loss $\ell_t(a_t)$ is the probability of a mistake, or as a real-valued prediction, where the loss is the absolute difference between the prediction and the outcome.

Prediction with Expert Advice

The minimax regret is the value of the game:

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

An easier game

Suppose that the adversary is constrained to choose the sequence ℓ_t so that some expert incurs no loss, that is,

$$\min_{a \in \mathcal{C}} \sum_{t=1}^{n} \ell_t(a) = 0.$$

How should we predict?

Prediction with Expert Advice

Halving Algorithm

[Littlestone, 1988]

Prediction with Expert Advice

Halving Algorithm

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

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

[Littlestone, 1988]

Prediction with Expert Advice

Halving Algorithm

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

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

- Choose

$$\hat{y}_t = a_t(f_t)^\top f_t = \text{majority}\left(\{f_t(j) : j \in C_t\}\right)$$

[Littlestone, 1988]

Prediction with Expert Advice

Halving Algorithm

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

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

- Choose

$$\hat{y}_t = a_t(f_t)^\top f_t = \text{majority}\left(\{f_t(j) : j \in C_t\}\right)$$

Theorem

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

[Littlestone, 1988]

Proof

If the strategy 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:

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

Proof

If the strategy 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).

Proof

If the strategy 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,

$$\sum_{t=1}^{n} \ell_t(a_t) \leq \log_2 \frac{|C_1|}{|C_{n+1}|}$$

Prediction with Expert Advice

Proof

If the strategy 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,

$$\sum_{t=1}^{n} \ell_t(a_t) \leq \log_2 \frac{|C_1|}{|C_{n+1}|} = \log_2 m - \log_2 |C_{n+1}|$$

Prediction with Expert Advice

Proof

If the strategy 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,

$$\sum_{t=1}^{n} \ell_t(a_t) \leq \log_2 \frac{|C_1|}{|C_{n+1}|} = \log_2 m - \log_2 |C_{n+1}| \leq \log_2 m.$$

Prediction with Expert Advice

We can do better with a randomized voting strategy.

[Karlin and Peres, 2016]

Prediction with Expert Advice

We can do better with a randomized voting strategy.

Random Leader

Choose $a_t(f_t)$ uniformly on

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

[Karlin and Peres, 2016]

Prediction with Expert Advice

We can do better with a randomized voting strategy.

Random Leader

Choose $a_t(f_t)$ uniformly on

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

Theorem

This strategy has regret no more than $H_m - 1$, where

$$H_m = \sum_{i=1}^{m} \frac{1}{i} \in (\ln m, \ln m + 1).$$

[Karlin and Peres, 2016]

Proof

We show that, at time t, the strategy can make no more than $H_{|C_t|} - 1$ mistakes from that time on.

Prediction with Expert Advice

Proof

We show that, at time t, the strategy can make no more than $H_{|C_t|} - 1$ mistakes from that time on.

- This is clearly true when $|C_t| = 1$: the strategy never makes another mistake.

Prediction with Expert Advice

Proof

We show that, at time t, the strategy can make no more than $H_{|C_t|} - 1$ mistakes from that time on.

- This is clearly true when $|C_t| = 1$: the strategy never makes another mistake.
- Suppose it is true for $|C_{t+1}| < k$, suppose that $|C_t| = k$, and suppose that j experts in C_t make a mistake at time t, where $1 \leq j \leq k - 1$.

Prediction with Expert Advice

Proof

We show that, at time t, the strategy can make no more than $H_{|C_t|} - 1$ mistakes from that time on.

- This is clearly true when $|C_t| = 1$: the strategy never makes another mistake.
- Suppose it is true for $|C_{t+1}| < k$, suppose that $|C_t| = k$, and suppose that j experts in C_t make a mistake at time t, where $1 \leq j \leq k - 1$. Then the expected number of mistakes made from time t onwards is no more than

$$\frac{j}{k} + H_{k-j} - 1$$

Prediction with Expert Advice

Proof

We show that, at time t, the strategy can make no more than $H_{|C_t|} - 1$ mistakes from that time on.

- This is clearly true when $|C_t| = 1$: the strategy never makes another mistake.
- Suppose it is true for $|C_{t+1}| < k$, suppose that $|C_t| = k$, and suppose that j experts in C_t make a mistake at time t, where $1 \leq j \leq k - 1$. Then the expected number of mistakes made from time t onwards is no more than
$$\frac{j}{k} + H_{k-j} - 1 \leq H_k - 1.$$

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Lower bound

Set $k = \lfloor \log_2 m \rfloor$ so that $2^k \leq m \leq 2^{k+1}$.

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Lower bound

Set $k = \lfloor \log_2 m \rfloor$ so that $2^k \leq m \leq 2^{k+1}$. Consider the following adversary strategy:

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Lower bound

Set $k = \lfloor \log_2 m \rfloor$ so that $2^k \leq m \leq 2^{k+1}$. Consider the following adversary strategy:

- Choose C_0 as the first k experts.

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Lower bound

Set $k = \lfloor \log_2 m \rfloor$ so that $2^k \leq m \leq 2^{k+1}$. Consider the following adversary strategy:

- Choose C_0 as the first k experts.
- At round $1 \leq t \leq k$, choose $C_{t+1} \subset C_t$ uniformly at random from subsets of size $|C_t|/2$.

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Lower bound

Set $k = \lfloor \log_2 m \rfloor$ so that $2^k \leq m \leq 2^{k+1}$. Consider the following adversary strategy:

- Choose C_0 as the first k experts.
- At round $1 \leq t \leq k$, choose $C_{t+1} \subset C_t$ uniformly at random from subsets of size $|C_t|/2$.
- Choose $y_t \in \{0, 1\}$ uniformly at random.

Prediction with Expert Advice

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

[Karlin and Peres, 2016]

Lower bound

Set $k = \lfloor \log_2 m \rfloor$ so that $2^k \leq m \leq 2^{k+1}$. Consider the following adversary strategy:

- Choose C_0 as the first k experts.
- At round $1 \leq t \leq k$, choose $C_{t+1} \subset C_t$ uniformly at random from subsets of size $|C_t|/2$.
- Choose $y_t \in \{0, 1\}$ uniformly at random.
- Set

$$f_t^i = \begin{cases} y_t & \text{for } i \in C_{t+1}, \\ 1 - y_t & \text{otherwise.} \end{cases}$$

Lower bound

Clearly, after k rounds there is still a perfect expert.

Lower bound

Clearly, after k rounds there is still a perfect expert. The expected number of mistakes of any player strategy is

$$\frac{k}{2}$$

Lower bound

Clearly, after k rounds there is still a perfect expert. The expected number of mistakes of any player strategy is

$$\frac{k}{2} = \frac{\lfloor \log_2 m \rfloor}{2}$$

Lower bound

Clearly, after k rounds there is still a perfect expert. The expected number of mistakes of any player strategy is

$$\frac{k}{2} = \frac{\lfloor \log_2 m \rfloor}{2} \geq \lfloor \log_4 m \rfloor.$$

Prediction with Expert Advice

Minimax strategy

Set $a_t(f_t)^\top f_t = \phi(p_t)\hat{y} + (1 - \phi(p_t))(1 - \hat{y})$, where

Prediction with Expert Advice

Minimax strategy

Set $a_t(f_t)^\top f_t = \phi(p_t)\hat{y} + (1 - \phi(p_t))(1 - \hat{y})$, where

$$\hat{y} = \text{majority}\left(\{f_t(j) : j \in C_t\}\right),$$

Prediction with Expert Advice

Minimax strategy

Set $a_t(f_t)^\top f_t = \phi(p_t)\hat{y} + (1 - \phi(p_t))(1 - \hat{y})$, where

$$\hat{y} = \text{majority}\left(\{f_t(j) : j \in C_t\}\right),$$
$$p_t = \frac{1}{|C_t|}\left|\left\{i \in C_t : f_t(i) = \text{majority}\left(\{f_t(j) : j \in C_t\}\right)\right\}\right|,$$

Prediction with Expert Advice

Minimax strategy

Set $a_t(f_t)^\top f_t = \phi(p_t)\hat{y} + (1 - \phi(p_t))(1 - \hat{y})$, where

$$\hat{y} = \text{majority}\left(\{f_t(j) : j \in C_t\}\right),$$

$$p_t = \frac{1}{|C_t|}\left|\left\{i \in C_t : f_t(i) = \text{majority}\left(\{f_t(j) : j \in C_t\}\right)\right\}\right|,$$

$$\phi(p) = 1 + \log_4 p.$$

Prediction with Expert Advice

Minimax strategy

Set $a_t(f_t)^\top f_t = \phi(p_t)\hat{y} + (1 - \phi(p_t))(1 - \hat{y})$, where

$$\hat{y} = \text{majority}\left(\{f_t(j) : j \in C_t\}\right),$$
$$p_t = \frac{1}{|C_t|}\left|\left\{i \in C_t : f_t(i) = \text{majority}\left(\{f_t(j) : j \in C_t\}\right)\right\}\right|,$$
$$\phi(p) = 1 + \log_4 p.$$

That is, follow the majority with probability $\phi(p_t)$.

Prediction with Expert Advice

Minimax strategy

Set $a_t(f_t)^\top f_t = \phi(p_t)\hat{y} + (1 - \phi(p_t))(1 - \hat{y})$, where

$$\hat{y} = \text{majority}\left(\{f_t(j) : j \in C_t\}\right),$$
$$p_t = \frac{1}{|C_t|}\left|\left\{i \in C_t : f_t(i) = \text{majority}\left(\{f_t(j) : j \in C_t\}\right)\right\}\right|,$$
$$\phi(p) = 1 + \log_4 p.$$

That is, follow the majority with probability $\phi(p_t)$.

(NB: $\phi(p) = 1$ corresponds to the halving algorithm.
$\phi(p) = p$ corresponds to voting uniformly on C_t.)

Prediction with Expert Advice

Proof

We'd like an upper bound on the expected number of mistakes of the form $\log_a m$.

Proof

We'd like an upper bound on the expected number of mistakes of the form $\log_a m$. To make the inductive proof of this bound work, we need to consider two cases.

Proof

We'd like an upper bound on the expected number of mistakes of the form $\log_a m$. To make the inductive proof of this bound work, we need to consider two cases. First, if the majority is correct ($y_t = \hat{y}$), then we need

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m.$$

Proof

We'd like an upper bound on the expected number of mistakes of the form $\log_a m$. To make the inductive proof of this bound work, we need to consider two cases. First, if the majority is correct ($y_t = \hat{y}$), then we need

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m.$$

Second, if the minority is correct, then we need

$$\log_a((1 - p_t)m) + \phi(p_t) \leq \log_a m.$$

Proof

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m,$$
$$\log_a((1 - p_t)m) + \phi(p_t) \leq \log_a m.$$

Prediction with Expert Advice

Proof

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m,$$
$$\log_a((1 - p_t)m) + \phi(p_t) \leq \log_a m.$$

Rearranging and combining, we need

$$1 + \log_a p_t \leq \phi(p_t) \leq -\log_a(1 - p_t)$$

Proof

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m,$$
$$\log_a((1 - p_t)m) + \phi(p_t) \leq \log_a m.$$

Rearranging and combining, we need

$$1 + \log_a p_t \leq \phi(p_t) \leq -\log_a(1 - p_t)$$
$$\Leftrightarrow \quad \log_a(ap_t) \leq \log_a\left(\frac{1}{1 - p_t}\right).$$

Prediction with Expert Advice

Proof

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m,$$
$$\log_a((1 - p_t)m) + \phi(p_t) \leq \log_a m.$$

Rearranging and combining, we need

$$1 + \log_a p_t \leq \phi(p_t) \leq -\log_a(1 - p_t)$$
$$\Leftrightarrow \quad \log_a(ap_t) \leq \log_a\left(\frac{1}{1 - p_t}\right).$$

The largest a satisfying $ap_t(1 - p_t) \leq 1$ is $a = 4$.

Proof

$$\log_a(p_t m) + (1 - \phi(p_t)) \leq \log_a m,$$
$$\log_a((1 - p_t)m) + \phi(p_t) \leq \log_a m.$$

Rearranging and combining, we need

$$1 + \log_a p_t \leq \phi(p_t) \leq -\log_a(1 - p_t)$$
$$\Leftrightarrow \quad \log_a(ap_t) \leq \log_a\left(\frac{1}{1 - p_t}\right).$$

The largest a satisfying $ap_t(1 - p_t) \leq 1$ is $a = 4$.
So any $\phi(p_t)$ between $1 + \log_4 p_t$ and $-\log_4(1 - p_t)$ will suffice.

Theorem

The minimax regret is between $\lfloor \log_4 m \rfloor$ and $\log_4 m$.

Online convex optimization

1. Binary prediction
   - With (perfect) expert advice
   - Minimax strategy
   - With imperfect experts: exponential weights
2. General online convex
3. Minimax strategies

Prediction with Expert Advice

We return to the voting setting, and allow even the best expert to make mistakes.

Prediction with Expert Advice

We return to the voting setting, and allow even the best expert to make mistakes.

Voting

The player votes for a mixture of experts:
we set $\mathcal{A} = \Delta^m$, the probability simplex on $\{1, \ldots, m\}$, and the loss function at time t is $\ell_t(a) = \sum_{i=1}^m a_i \ell_t(e_i)$, where $e_i \in \Delta^m$ is zero in all but the ith coordinate, and $\ell_t(e_i) \in \{0, 1\}$ is the indicator for the ith expert making an incorrect forecast at time t.

Prediction with Expert Advice

Exponential Weights

[Littlestone and Warmuth, 1994]

Prediction with Expert Advice

Exponential Weights

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

[Littlestone and Warmuth, 1994]

Exponential Weights

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

$$w_1^i = 1,$$

[Littlestone and Warmuth, 1994]

Prediction with Expert Advice

Exponential Weights

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

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

[Littlestone and Warmuth, 1994]

Prediction with Expert Advice

Exponential Weights

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

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

- Here, $\eta > 0$ is a parameter of the algorithm.

[Littlestone and Warmuth, 1994]

Prediction with Expert Advice

Exponential Weights

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

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

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

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

[Littlestone and Warmuth, 1994]

Prediction with Expert Advice

Theorem

The exponential weights strategy with parameter

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

has regret satisfying

$$R_n \leq \sqrt{\frac{n \ln m}{2}}.$$

[Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth, 1997]

Prediction with Expert Advice

Proof Idea

We use a measure of progress:

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

Proof Idea

We use a measure of progress:

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

1. W_n grows at least as

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

Prediction with Expert Advice

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

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

Prediction with Expert Advice

Proof idea:

$$\ln \frac{W_{n+1}}{W_1}$$

Prediction with Expert Advice

Proof idea:

$$\ln \frac{W_{n+1}}{W_1} = \ln \left(\sum_{i=1}^{m} w_{n+1}^{i} \right) - \ln m$$

Prediction with Expert Advice

Proof idea:

$$\ln \frac{W_{n+1}}{W_1} = \ln\left(\sum_{i=1}^{m} w_{n+1}^{i}\right) - \ln m$$
$$= \ln\left(\sum_{i=1}^{m} \exp\left(-\eta \sum_{t} \ell_t(e_i)\right)\right) - \ln m$$

Prediction with Expert Advice

Proof idea:

$$\ln \frac{W_{n+1}}{W_1} = \ln \left(\sum_{i=1}^{m} w_{n+1}^{i} \right) - \ln m$$

$$= \ln \left(\sum_{i=1}^{m} \exp \left(-\eta \sum_{t} \ell_t(e_i) \right) \right) - \ln m$$

$$\geq \ln \left(\max_i \exp \left(-\eta \sum_{t} \ell_t(e_i) \right) \right) - \ln m$$

Prediction with Expert Advice

Proof idea:

$$
\begin{aligned}
\ln \frac{W_{n+1}}{W_1} &= \ln \left(\sum_{i=1}^{m} w_{n+1}^{i} \right) - \ln m \\
&= \ln \left(\sum_{i=1}^{m} \exp \left(-\eta \sum_{t} \ell_t(e_i) \right) \right) - \ln m \\
&\geq \ln \left(\max_{i} \exp \left(-\eta \sum_{t} \ell_t(e_i) \right) \right) - \ln m \\
&= -\eta \min_{i} \left(\sum_{t} \ell_t(e_i) \right) - \ln m
\end{aligned}
$$

Prediction with Expert Advice

Proof idea:

$$
\begin{aligned}
\ln \frac{W_{n+1}}{W_1} &= \ln \left(\sum_{i=1}^{m} w_{n+1}^{i} \right) - \ln m \\
&= \ln \left(\sum_{i=1}^{m} \exp \left(-\eta \sum_{t} \ell_t(e_i) \right) \right) - \ln m \\
&\geq \ln \left(\max_{i} \exp \left(-\eta \sum_{t} \ell_t(e_i) \right) \right) - \ln m \\
&= -\eta \min_{i} \left(\sum_{t} \ell_t(e_i) \right) - \ln m \\
&= -\eta \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) - \ln m.
\end{aligned}
$$

Prediction with Expert Advice

Proof idea:

$$\ln \frac{W_{t+1}}{W_t}$$

Prediction with Expert Advice

Proof idea:

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

Prediction with Expert Advice

Proof idea:

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

Proof idea:

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

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

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

Prediction with Expert Advice

Proof idea:

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

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

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

Prediction with Expert Advice

Proof idea:

$$-\eta \inf_{a\in\mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) - \ln m \leq \ln \frac{W_{n+1}}{W_1} \leq -\eta \sum_{t=1}^{n} \ell_t(a_t) + \frac{n\eta^2}{8}.$$

Prediction with Expert Advice

Proof idea:

$$-\eta \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) - \ln m \leq \ln \frac{W_{n+1}}{W_1} \leq -\eta \sum_{t=1}^{n} \ell_t(a_t) + \frac{n\eta^2}{8}.$$

Thus,

$$R_n \leq \frac{\ln m}{\eta} + \frac{\eta n}{8}.$$

Prediction with Expert Advice

Proof idea:

$$-\eta \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) - \ln m \leq \ln \frac{W_{n+1}}{W_1} \leq -\eta \sum_{t=1}^{n} \ell_t(a_t) + \frac{n\eta^2}{8}.$$

Thus,

$$R_n \leq \frac{\ln m}{\eta} + \frac{\eta n}{8}.$$

Choosing the optimal η gives the result:

Prediction with Expert Advice

Proof idea:

$$-\eta \inf_{a\in\mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) - \ln m \leq \ln \frac{W_{n+1}}{W_1} \leq -\eta \sum_{t=1}^{n} \ell_t(a_t) + \frac{n\eta^2}{8}.$$

Thus,

$$R_n \leq \frac{\ln m}{\eta} + \frac{\eta n}{8}.$$

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 $\sqrt{\frac{n \ln m}{2}}$.

Prediction with Expert Advice

Key Points

For a finite set of actions (experts):

Prediction with Expert Advice

Key Points

For a finite set of actions (experts):

- If one action is perfect (i.e., has zero loss), the minimax strategy gives per round regret of
$$\frac{\log_4 m}{n}.$$

Prediction with Expert Advice

Key Points

For a finite set of actions (experts):

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

$$\frac{\log_4 m}{n}.$$

- Exponential weights gives per round regret of

$$\sqrt{\frac{\ln m}{2n}}.$$

Prediction with Expert Advice

In the proof, the only properties of ℓ_t that we used were

Prediction with Expert Advice

In the proof, the only properties of ℓ_t that we used were

1. boundedness: $\ell_t(e_i) \in [0, 1]$ (for Hoeffding's inequality), and

Prediction with Expert Advice

In the proof, the only properties of ℓ_t that we used were

1. boundedness: $\ell_t(e_i) \in [0, 1]$ (for Hoeffding's inequality), and
2. linearity,

$$\ell_t(a_t) = \sum_i \ell_t(e_i) w_t^i.$$

Prediction with Expert Advice

In the proof, the only properties of ℓ_t that we used were

1. boundedness: $\ell_t(e_i) \in [0, 1]$ (for Hoeffding's inequality), and
2. linearity,

$$\ell_t(a_t) = \sum_i \ell_t(e_i) w_t^i.$$

For linearity, an inequality would have sufficed,

$$\ell_t(a_t) \leq \sum_i \ell_t(e_i) w_t^i,$$

Prediction with Expert Advice

In the proof, the only properties of ℓ_t that we used were

1. boundedness: $\ell_t(e_i) \in [0, 1]$ (for Hoeffding's inequality), and
2. linearity,

$$\ell_t(a_t) = \sum_i \ell_t(e_i) w_t^i.$$

For linearity, an inequality would have sufficed,

$$\ell_t(a_t) \leq \sum_i \ell_t(e_i) w_t^i,$$

which corresponds to convexity of ℓ_t.

Online convex optimization

1. Binary prediction
   - With (perfect) expert advice
   - Minimax strategy
   - With imperfect experts: exponential weights
2. General online convex
   - Empirical minimization fails
   - Gradient algorithm.
   - A regularization viewpoint
   - Bregman divergence
   - Properties of regularization
   - Linearization
   - Mirror descent
   - Regret bounds
   - Strongly convex losses
   - Adaptive regularization
3. Minimax strategies

Online convex optimization

The problem

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

Online convex optimization

The problem

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

Minimax regret

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

Online convex optimization

Empirical minimization fails

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')

Online convex optimization

Empirical minimization fails

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 $\mathcal{A} = [-1, 1]$, $\mathcal{L} = \{a \mapsto v \cdot a : |v| \leq 1\}$.

Online convex optimization

Empirical minimization fails

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:

$$a_1 = 0,$$

Empirical minimization fails

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:

$$\ell_1(a) = \frac{1}{2}a,$$
$$a_1 = 0,$$

Online convex optimization

Empirical minimization fails

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

$$\ell_1(a) = \frac{1}{2} a,$$

$$a_1 = 0, \qquad a_2 = -1,$$

Empirical minimization fails

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:

$$\begin{aligned} \ell_1(a) &= \frac{1}{2}a, & \ell_2(a) &= -a, \\ a_1 &= 0, & a_2 &= -1, \end{aligned}$$

Online convex optimization

Empirical minimization fails

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:

$$\ell_1(a) = \frac{1}{2}a, \quad \ell_2(a) = -a,$$
$$a_1 = 0, \qquad a_2 = -1, \qquad a_3 = 1,$$

Online convex optimization

Empirical minimization fails

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

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

Online convex optimization

Empirical minimization fails

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

$$\ell_1(a) = \frac{1}{2}a, \quad \ell_2(a) = -a, \quad \ell_3(a) = a,$$
$$a_1 = 0, \qquad a_2 = -1, \qquad a_3 = 1, \qquad a_4 = -1,$$

Online convex optimization

Empirical minimization fails

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:

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

Online convex optimization

Empirical minimization fails

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:

$$\ell_1(a) = \frac{1}{2}a, \quad \ell_2(a) = -a, \quad \ell_3(a) = a, \quad \ell_4(a) = -a,$$
$$a_1 = 0, \qquad a_2 = -1, \qquad a_3 = 1, \qquad a_4 = -1, \qquad a_5 = 1,$$

Online convex optimization

Empirical minimization fails

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:

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

Online convex optimization

Empirical minimization fails

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:

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

- $a^* = 0$ shows $\min_{a \in \mathbb{R}^d} \sum_{t=1}^n \ell_t(a) \leq 0$, but $\sum_{t=1}^n \ell_t(a_t) = n - 1$.

Online convex optimization

- Choosing a_t to minimize past losses can fail.

Online convex optimization

- Choosing a_t to minimize past losses can fail.
- The strategy must avoid overfitting, just as in probabilistic settings.

Online convex optimization

- 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.

Online convex optimization

- 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

Online convex optimization: Gradient Method

$a_1 \in \mathcal{A}$,

[Zinkevich, 2003]

Online convex optimization: Gradient Method

$$a_1 \in \mathcal{A}, \qquad a_{t+1} = \qquad (a_t - \eta \nabla \ell_t(a_t)),$$

[Zinkevich, 2003]

Online convex optimization: Gradient Method

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

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

[Zinkevich, 2003]

Online convex optimization: Gradient Method

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

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

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

[Zinkevich, 2003]

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

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

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

Theorem

the gradient strategy with $\eta = D/(G\sqrt{n})$

[Zinkevich, 2003]

Online convex optimization: Gradient Method

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

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 $G = \max_t \|\nabla \ell_t(a_t)\|$
the gradient strategy with $\eta = D/(G\sqrt{n})$

[Zinkevich, 2003]

Online convex optimization: Gradient Method

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

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 $D = \operatorname{diam}(\mathcal{A})$,
the gradient strategy with $\eta = D/(G\sqrt{n})$

[Zinkevich, 2003]

Online convex optimization: Gradient Method

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

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 $D = \mathrm{diam}(\mathcal{A})$,
the gradient strategy with $\eta = D/(G\sqrt{n})$ has regret satisfying

$$R_n \leq GD\sqrt{n}.$$

[Zinkevich, 2003]

Example

$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq 1\}$,

Online convex optimization: Gradient Method

Example

$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq 1\}$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\| \leq 1\}$.

Online convex optimization: Gradient Method

Example

$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq 1\}$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\| \leq 1\}$.
$D = 2$,

Online convex optimization: Gradient Method

Example

$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq 1\}, \;\; \mathcal{L} = \{a \mapsto v \cdot a : \|v\| \leq 1\}.$
$D = 2, \;\; G \leq 1.$

Online convex optimization: Gradient Method

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 $2\sqrt{n}$.

Online convex optimization: Gradient Method

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

Example

$\mathcal{A} = \Delta^m$,

Online convex optimization: Gradient Method

Example

$\mathcal{A} = \Delta^m, \;\; \mathcal{L} = \{a \mapsto v \cdot a : \|v\|_\infty \leq 1\}.$

Online convex optimization: Gradient Method

Example

$\mathcal{A} = \Delta^m$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\|_\infty \leq 1\}$.
$D = \sqrt{2}$,

Online convex optimization: Gradient Method

Example

$\mathcal{A} = \Delta^m, \;\; \mathcal{L} = \{a \mapsto v \cdot a : \|v\|_\infty \leq 1\}.$
$D = \sqrt{2}, \;\; G \leq \sqrt{m}.$

Online convex optimization: Gradient Method

Example

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

Online convex optimization: Gradient Method

Example

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

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

Online convex optimization: Gradient Method

Proof:

Define $$\tilde{a}_{t+1} = a_t - \eta \nabla \ell_t(a_t),$$

Proof:

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

Online convex optimization: Gradient Method

Proof:

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

Fix a comparator $a \in \mathcal{A}$

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

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

$$\|a_{t+1} - a\|^2$$

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

$$\|a_{t+1} - a\|^2 \leq \|\tilde{a}_{t+1} - a\|^2$$

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

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

Online convex optimization: Gradient Method

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a))$$

Online convex optimization: Gradient Method

By convexity,

$$\sum_{t=1}^{n}(\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} \nabla\ell_t(a_t) \cdot (a_t - a)$$

Online convex optimization: Gradient Method

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: Gradient Method

By convexity,

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

Online Convex Optimization

1. Binary prediction
2. General online convex
   - Empirical minimization fails
   - Gradient algorithm
   - A regularization viewpoint
   - Bregman divergence
   - Properties of regularization
   - Linearization
   - Mirror descent
   - Regret bounds
   - Strongly convex losses
   - Adaptive regularization
3. Minimax strategies

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

- Suppose ℓ_t is linear: $\ell_t(a) = g_t \cdot a$.

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

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

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

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

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

- 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 $a_{t+1} = a_t - \eta \nabla \ell_t(a_t)$.
- Indeed, setting the derivative to zero gives

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

- 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 $a_{t+1} = a_t - \eta \nabla \ell_t(a_t)$.
- Indeed, setting the derivative to zero gives

$$a_{t+1} = -\eta \sum_{s=1}^{t} \nabla \ell_s$$

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

- 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 $a_{t+1} = a_t - \eta \nabla \ell_t(a_t)$.
- Indeed, setting the derivative to zero gives

$$a_{t+1} = -\eta \sum_{s=1}^{t} \nabla \ell_s$$

$$a_t = -\eta \sum_{s=1}^{t-1} \nabla \ell_s.$$

A Regularization Viewpoint

An observation: gradient algorithm is regularized minimization.

- 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 $a_{t+1} = a_t - \eta \nabla \ell_t(a_t)$.
- Indeed, setting the derivative to zero gives

$$a_{t+1} = -\eta \sum_{s=1}^{t} \nabla \ell_s = a_t - \eta \nabla \ell_t.$$

$$a_t = -\eta \sum_{s=1}^{t-1} \nabla \ell_s.$$

Online Convex Optimization: Regularization

Definition: Regularized minimization

Consider the family of strategies of the form:

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

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

Online Convex Optimization: Regularization

Regularized minimization

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

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.

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.

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.

Online Convex Optimization: Regularization

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.

Online Convex Optimization: Regularization

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 $D_{\Phi_{t-1}}$:

Online Convex Optimization: Regularization

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 $D_{\Phi_{t-1}}$:

Definition

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

Online Convex Optimization: Regularization

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 $D_{\Phi_{t-1}}$:

Definition

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

So

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

Online Convex Optimization: Regularization

Definition: Bregman Divergence

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

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

[Bregman, 1967]

Online Convex Optimization: Regularization

Definition: Bregman Divergence

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, 1967]

Bregman Divergence

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

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

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

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

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

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

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

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

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

$$D_\Phi(a, b) = \sum_i (a_i(\ln a_i - 1) - b_i(\ln b_i - 1) - \ln b_i(a_i - b_i))$$

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.

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:

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,

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 $\mathcal{A}$, $\|\nabla\Phi(a_n)\| \to \infty$.

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

Bregman Divergence

Properties

1. $D_\Phi \geq 0$

Bregman Divergence

Properties

1. $D_\Phi \geq 0, \; D_\Phi(a, a) = 0.$

Bregman Divergence

Properties

1. $D_\Phi \geq 0, \ D_\Phi(a, a) = 0.$
2. $D_{A+B} = D_A + D_B.$

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 $\mathcal{A}$.

Bregman Divergence

Properties

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

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

Bregman Divergence

Properties

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

Bregman Divergence

Properties

Bregman Divergence

Properties

6. For ℓ affine, $D_{\Phi+\ell} = D_{\Phi}$.

Bregman Divergence

Properties

6. For ℓ affine, $D_{\Phi+\ell} = D_{\Phi}$.
7. For Φ^* the Legendre dual of Φ,

$$\nabla\Phi^* = (\nabla\Phi)^{-1},$$

Bregman Divergence

Properties

6. For ℓ affine, $D_{\Phi+\ell} = D_{\Phi}$.
7. For Φ^* the Legendre dual of Φ,

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

Bregman Divergence

Definition: Legendre Dual

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

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

Definition: Legendre Dual

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

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

- Φ^* is Legendre.

Bregman Divergence

Definition: Legendre Dual

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

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

- Φ^* is Legendre.
- $\mathrm{dom}(\Phi^*) = \nabla \Phi(\mathrm{int\ dom}\ \Phi)$.

Bregman Divergence

Definition: Legendre Dual

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

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

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

Bregman Divergence

Definition: Legendre Dual

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

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

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

Bregman Divergence

Definition: Legendre Dual

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

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

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

Bregman Divergence

Example

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

Bregman Divergence

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

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

Bregman Divergence

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

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

Bregman Divergence

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

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

so

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

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

Online Convex Optimization

1. Binary prediction
2. General online convex
    - Empirical minimization fails
    - Gradient algorithm
    - A regularization viewpoint
    - Bregman divergence
    - Properties of regularization
    - Linearization
    - Mirror descent
    - Regret bounds
    - Strongly convex losses
    - Adaptive regularization
3. Minimax strategies

Properties of Regularization Methods

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

Theorem

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

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

Properties of Regularization Methods

In the unconstrained case ($\mathcal{A} = \mathbb{R}^d$), regularized minimization is equivalent to minimizing the latest loss plus 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, $\left(\Phi_t(a) := \eta \sum_{s=1}^{t} \ell_s(a) + R(a)\right)$

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

Properties of Regularization Methods

Proof:

By the definition of Φ_t, $\left(\Phi_t(a) := \eta \sum_{s=1}^{t} \ell_s(a) + R(a)\right)$

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

The derivative wrt a is

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

Properties of Regularization Methods

Proof:

By the definition of Φ_t, $\quad (\Phi_t(a) := \eta \sum_{s=1}^{t} \ell_s(a) + R(a))$

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

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

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

Properties of Regularization Methods

Proof:

By the definition of Φ_t, $\qquad (\Phi_t(a) := \eta \sum_{s=1}^{t} \ell_s(a) + R(a))$

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

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:

Properties of Regularization Methods

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

Theorem

For

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \Phi_t(a),$$

Properties of Regularization Methods

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

Theorem

For

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

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

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

Proof:

Let a'_{t+1} denote $\Pi_{\mathcal{A}}^{\Phi_t}(\tilde{a}_{t+1})$.

Proof:

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

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

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,

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

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

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

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},

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

Conversely,

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

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

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:

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:

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

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:

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

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

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:

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

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

(e.g., R squared Euclidean norm)

Online Convex Optimization

1. Binary prediction
2. General online convex
    - Empirical minimization fails
    - Gradient algorithm
    - A regularization viewpoint
    - Bregman divergence
    - Properties of regularization
    - Linearization
    - Mirror descent
    - Regret bounds
    - Strongly convex losses
    - Adaptive regularization
3. Minimax strategies

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.

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

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

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

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

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

Key Point:

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

Properties of Regularization Methods

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.

Online convex optimization

1. Binary prediction
2. General online convex
   - Empirical minimization fails
   - Gradient algorithm
   - A regularization viewpoint
   - Bregman divergence
   - Properties of regularization
   - Linearization
   - Mirror descent
   - Regret bounds
   - Strongly convex losses
   - Adaptive regularization
3. Minimax strategies

Regularization Methods: Mirror Descent

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

see [Nemirovsky and Yudin, 1983]

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

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

see [Nemirovsky and Yudin, 1983]

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

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

see [Nemirovsky and Yudin, 1983]

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 $\tilde{a}_{t+1}$.

see [Nemirovsky and Yudin, 1983]

Regularization Methods: Mirror Descent

Proof:

For the unconstrained minimization, we have

Regularization Methods: Mirror Descent

Proof:

For the unconstrained minimization, we have

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

Regularization Methods: Mirror Descent

Proof:

For the unconstrained minimization, we have

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

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

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

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. Binary prediction
2. General online convex
   - Empirical minimization fails
   - Gradient algorithm
   - A regularization viewpoint
   - Bregman divergence
   - Properties of regularization
   - Linearization
   - Mirror descent
   - Regret bounds
   - Strongly convex losses
   - Adaptive regularization
3. Minimax strategies

Regularization methods: Regret bounds

Recall: 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

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

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

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

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

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

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

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

Regularization methods: Regret

Theorem

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

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

Notice that, because a_{t+1} is the unconstrained minimizer of Φ_t,

$$D_{\Phi_t}(a_t, a_{t+1}) = D_{\Phi_t^*}(\nabla \Phi_t(a_{t+1}), \nabla \Phi_t(a_t))$$

Regularization methods: Regret

Theorem

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

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

Notice that, because a_{t+1} is the unconstrained minimizer of Φ_t,

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

Regularization methods: Regret

Theorem

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

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

Notice that, because a_{t+1} is the unconstrained minimizer of Φ_t,

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

Regularization methods: Regret

Theorem

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

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

Notice that, because a_{t+1} is the unconstrained minimizer of Φ_t,

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

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

Example

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

Regularization methods: Regret

Example

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

$$\sum_{t=1}^{n} \ell_t(a_t) \leq \inf_{a\in\mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) + \frac{\|a^* - a_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{n} \|g_t\|^2.$$

Example

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

$$\sum_{t=1}^{n} \ell_t(a_t) \leq \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) + \frac{\|a^* - a_1\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|g_t\|^2.$$

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

Regularization methods: Regret

Seeing the future gives small regret

For regularized minimization, that is,

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

see also [Kalai and Vempala, 2005]

Regularization methods: Regret

Seeing the future gives small regret

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 $a \in \mathcal{A}$,

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

see also [Kalai and Vempala, 2005]

Regularization methods: Regret

Seeing the future gives small regret

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 $a \in \mathcal{A}$,

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

(NB: This is cheating!)

see also [Kalai and Vempala, 2005]

Regularization methods: Regret

Proof:

Regularization methods: Regret

Proof:

Since a_{t+1} minimizes Φ_t,

Regularization methods: Regret

Proof:

Since a_{t+1} minimizes Φ_t,

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

Regularization methods: Regret

Proof:

Since a_{t+1} minimizes Φ_t,

$$\begin{aligned}\eta \sum_{s=1}^{t} \ell_s(a) + R(a) &\geq \eta \sum_{s=1}^{t} \ell_s(a_{t+1}) + R(a_{t+1}) \\ &= \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_{t+1}) + R(a_{t+1})\end{aligned}$$

Regularization methods: Regret

Proof:

Since a_{t+1} minimizes Φ_t, and a_t minimizes Φ_{t-1},

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

Regularization methods: Regret

Proof:

Since a_{t+1} minimizes Φ_t, and a_t minimizes Φ_{t-1},

$$\begin{aligned}
\eta \sum_{s=1}^{t} \ell_s(a) + R(a) &\geq \eta \sum_{s=1}^{t} \ell_s(a_{t+1}) + R(a_{t+1}) \\
&= \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_{t+1}) + R(a_{t+1}) \\
&\geq \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_t) + R(a_t) \\
&\vdots \\
&\geq \eta \sum_{s=1}^{t} \ell_s(a_{s+1}) + R(a_1).
\end{aligned}$$

Regularization methods: Regret

Theorem

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

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

Regularization methods: Regret

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 $a \in \mathcal{A}$,

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

Regularization methods: Regret

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 $\ell_t(a_t) - \ell_t(a_{t+1})$?

Regularization methods: Regret

Definition

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

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

Regularization methods: Regret

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

Regularization methods: Regret

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 a_{t+1} minimizes Φ_t and $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$:

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

Regularization methods: Regret

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 a_{t+1} minimizes Φ_t and $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$:

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

Note that the definition implies a generalization of the Cauchy-Schwarz inequality: for $\|a\| > 0$,

$$v \cdot \frac{a}{\|a\|} \leq \|v\|_*.$$

Proof:

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

Proof:

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

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

Regularization methods: Regret

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,

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

Regularization methods: Regret

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

This leads to the regret bound:

Corollary

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

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

Regularization methods: Regret

This leads to the regret bound:

Corollary

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

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

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

Regularization methods: Regret

Example

Consider $R(a) = \frac{1}{2}\|a\|^2$, $a_1 = 0$, and $\mathcal{A}$ contained in a Euclidean ball of diameter D.

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.

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

Example

Consider $\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$.

Regularization methods: Regret

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).

Regularization methods: Regret

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,

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

Regularization methods: Regret

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 $\| \cdot \|_1$.

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 $\|\cdot\|_1$. Also, $R(a) - R(a_1) \leq \ln m$.

Regularization methods: Regret

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 $\|\cdot\|_1$. Also, $R(a) - R(a_1) \leq \ln m$.
Suppose that $\|g_t\|_\infty \leq 1$.

Regularization methods: Regret

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

Regularization methods: Regret

Example

$\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$.

Regularization methods: Regret

Example

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

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

Regularization methods: Regret

Example

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

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

Regularization methods: Regret

Example

$\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$.

What are the updates?

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

Regularization methods: Regret

Example

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

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

Regularization methods: Regret

Example

$\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$.

What are the updates?

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

where the ln and exp functions are applied component-wise.

Regularization methods: Regret

Example

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

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

where the ln and exp functions are applied component-wise.
This is exponentiated gradient: mirror descent with $\nabla R = \ln$.

Regularization methods: Regret

Example

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

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

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, $\Pi_{\mathcal{A}}^{R}(\tilde{a}) = \tilde{a}/\|\tilde{a}\|_1$.

Regularization methods: Regret

Notice that when the losses are linear, exponentiated gradient is exactly the exponential weights strategy we discussed for a finite comparison class.

Regularization methods: Regret

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 $O(\sqrt{mn})$.

Regularization methods: Extensions

- Instead of

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

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

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

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.

Online convex optimization

1. Binary prediction
2. General online convex
    - Empirical minimization fails
    - Gradient algorithm
    - A regularization viewpoint
    - Bregman divergence
    - Properties of regularization
    - Linearization
    - Mirror descent
    - Regret bounds
    - Strongly convex losses
    - Adaptive regularization
3. Minimax strategies

Regularization methods: 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 ℓ_t, it is $O(\log n)$, versus $O(\sqrt{n})$.

Regularization methods: Strongly convex losses

Some intuition about time-varying η:

Consider

$$\Phi_t(a) = \sum_{s=1}^{t} \eta_s \ell_s(a) + R(a), \qquad a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a).$$

Regularization methods: Strongly convex losses

Some intuition about time-varying η:

Consider

$$\Phi_t(a) = \sum_{s=1}^{t} \eta_s \ell_s(a) + R(a), \qquad a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a).$$

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

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

Regularization methods: Strongly convex losses

Some intuition about time-varying η:

Consider

$$\Phi_t(a) = \sum_{s=1}^{t} \eta_s \ell_s(a) + R(a), \qquad a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a).$$

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

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

(Easy to check. Use $\nabla\Phi_t(a_{t+1}) = \nabla\Phi_{t-1}(a_t) = 0$.)

Regularization methods: Strongly convex losses

Some intuition about time-varying η:

Consider

$$\Phi_t(a) = \sum_{s=1}^{t} \eta_s \ell_s(a) + R(a), \qquad a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a).$$

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

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

(Easy to check. Use $\nabla \Phi_t(a_{t+1}) = \nabla \Phi_{t-1}(a_t) = 0$.)
What keeps the last two terms small?

Regularization methods: Strongly convex losses

Some intuition about time-varying η:

Consider

$$\Phi_t(a) = \sum_{s=1}^{t} \eta_s \ell_s(a) + R(a), \qquad a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a).$$

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

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

(Easy to check. Use $\nabla \Phi_t(a_{t+1}) = \nabla \Phi_{t-1}(a_t) = 0$.)
What keeps the last two terms small? If we linearize the ℓ_t, we have

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

Regularization methods: Strongly convex losses

Some intuition about time-varying η:

Consider

$$\Phi_t(a) = \sum_{s=1}^{t} \eta_s \ell_s(a) + R(a), \qquad a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a).$$

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

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

(Easy to check. Use $\nabla\Phi_t(a_{t+1}) = \nabla\Phi_{t-1}(a_t) = 0$.)
What keeps the last two terms small? If we linearize the ℓ_t, we have

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

which requires $\eta_t \approx$ constant. 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 $a, b \in \mathbb{R}^d$,

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

[B., Hazan, Rakhlin, 2007]

Regularization methods: Strongly convex losses

Theorem

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

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

and R is strongly convex wrt $\|\cdot\|$,

[B., Hazan, Rakhlin, 2007]

Regularization methods: Strongly convex losses

Theorem

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

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

and R is strongly convex wrt $\| \cdot \|$, then for any $a \in \mathcal{A}$, mirror descent,

$$a_{t+1} = \Pi_{\mathcal{A}}^R \left((\nabla R)^{-1} \left(\nabla R(a_t) - \eta_t \nabla \ell_t(a_t) \right) \right)$$

with $\eta_t \geq \frac{2}{t\sigma}$ has regret

$$\sum_{t=1}^n \ell_t(a_t) - \sum_{t=1}^n \ell_t(a) \leq \sum_{t=1}^n \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) \leq \sum_{t=1}^n \eta_t \|\nabla \ell_t(a_t)\|_*^2.$$

[B., Hazan, Rakhlin, 2007]

Regularization methods: Strongly convex losses

Proof

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right).$$

Regularization methods: Strongly convex losses

Proof

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right).$$

Define: $\tilde{a}_{t+1}$ so that $a_{t+1} = \Pi_{\mathcal{A}}^{R}(\tilde{a}_{t+1})$:

Regularization methods: Strongly convex losses

Proof

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right).$$

Define: $\tilde{a}_{t+1}$ so that $a_{t+1} = \Pi_{\mathcal{A}}^{R}(\tilde{a}_{t+1})$:

$$\tilde{a}_{t+1} := \nabla R^{-1} (\nabla R(a_t) - \eta_t \nabla \ell_t(a_t)),$$

Regularization methods: Strongly convex losses

Proof

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right).$$

Define: $\tilde{a}_{t+1}$ so that $a_{t+1} = \Pi_{\mathcal{A}}^{R}(\tilde{a}_{t+1})$:

$$\tilde{a}_{t+1} := \nabla R^{-1} (\nabla R(a_t) - \eta_t \nabla \ell_t(a_t)),$$

and hence

$$\nabla R^{-1} (\tilde{a}_{t+1}) := \nabla R(a_t) - \eta_t \nabla \ell_t(a_t).$$

Proof

$$\begin{aligned}
&\nabla \ell_t(a_t) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} \left(\nabla R(a_t) - \nabla R(\tilde{a}_{t+1}) \right) \cdot (a_t - a)
\end{aligned}$$

Regularization methods: Strongly convex losses

Proof

$$\begin{aligned}
&\nabla \ell_t(a_t) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - a)
\end{aligned}$$

where the first equality follows from the definition of $\tilde{a}_{t+1}$,

Regularization methods: Strongly convex losses

Proof

$$\begin{aligned}
&\nabla \ell_t(a_t) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} (D_R(a, a_t) - D_R(a, \tilde{a}_{t+1}) + D_R(a_t, \tilde{a}_{t+1}))
\end{aligned}$$

where the first equality follows from the definition of $\tilde{a}_{t+1}$,

Proof

$$
\begin{aligned}
&\nabla \ell_t(a_t) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} \left(\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})\right) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} \left(D_R(a, a_t) - D_R(a, \tilde{a}_{t+1}) + D_R(a_t, \tilde{a}_{t+1})\right)
\end{aligned}
$$

where the first equality follows from the definition of $\tilde{a}_{t+1}$,
the second follows from the definition of Bregman divergence,

Regularization methods: Strongly convex losses

Proof

$$
\begin{aligned}
&\nabla \ell_t(a_t) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} (D_R(a, a_t) - D_R(a, \tilde{a}_{t+1}) + D_R(a_t, \tilde{a}_{t+1})) \\
&\leq \frac{1}{\eta_t} (D_R(a, a_t) - D_R(a, a_{t+1}) + D_R(a_t, \tilde{a}_{t+1})) ,
\end{aligned}
$$

where the first equality follows from the definition of $\tilde{a}_{t+1}$,
the second follows from the definition of Bregman divergence,

Regularization methods: Strongly convex losses

Proof

$$
\begin{aligned}
&\nabla \ell_t(a_t) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} \left(\nabla R(a_t) - \nabla R(\tilde{a}_{t+1}) \right) \cdot (a_t - a) \\
&= \frac{1}{\eta_t} \left(D_R(a, a_t) - D_R(a, \tilde{a}_{t+1}) + D_R(a_t, \tilde{a}_{t+1}) \right) \\
&\leq \frac{1}{\eta_t} \left(D_R(a, a_t) - D_R(a, a_{t+1}) + D_R(a_t, \tilde{a}_{t+1}) \right),
\end{aligned}
$$

where the first equality follows from the definition of $\tilde{a}_{t+1}$,
the second follows from the definition of Bregman divergence,
and the inequality follows from the Pythagorean Theorem for D_R (for $a^* = \Pi^{\Phi}_{\mathcal{A}}(b)$ and $a \in \mathcal{A}$, $D_\Phi(a, b) \geq D_\Phi(a, a^*) + D_\Phi(a^*, b)$.)

Regularization methods: Strongly convex losses

Proof

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a))$$
$$\leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t)\right)$$

Regularization methods: Strongly convex losses

Proof

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a))$$

$$\leq \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right)$$

$$\leq \sum_{t=1}^{n} \left(\frac{1}{\eta_t} (D_R(a, a_t) - D_R(a, a_{t+1}) + D_R(a_t, \tilde{a}_{t+1})) - \frac{\sigma}{2} D_R(a, a_t) \right)$$

Regularization methods: Strongly convex losses

Proof

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

Regularization methods: Strongly convex losses

Proof

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

And choosing $\eta_t = c/t$ for $c \ge 2/\sigma$ eliminates the second and third terms.

Proof

Also,

$$D_R(a_t, \tilde{a}_{t+1}) \leq D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t)$$

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned} D_R(a_t, \tilde{a}_{t+1}) &\leq D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\ &= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \end{aligned}$$

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned} D_R(a_t, \tilde{a}_{t+1}) &\leq D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\ &= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \\ &= \eta_t \nabla \ell_t(a_t) \cdot (a_t - \tilde{a}_{t+1}) \end{aligned}$$

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned} D_R(a_t, \tilde{a}_{t+1}) &\le D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\ &= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \\ &= \eta_t \nabla \ell_t(a_t) \cdot (a_t - \tilde{a}_{t+1}) \end{aligned}$$

where the second equality is from the definition of $\tilde{a}_{t+1}$

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned}
D_R(a_t, \tilde{a}_{t+1}) &\le D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\
&= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \\
&= \eta_t \nabla \ell_t(a_t) \cdot (a_t - \tilde{a}_{t+1}) \\
&\le \eta_t \|\nabla \ell_t(a_t)\|_* \|a_t - \tilde{a}_{t+1}\|
\end{aligned}$$

where the second equality is from the definition of $\tilde{a}_{t+1}$

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned} D_R(a_t, \tilde{a}_{t+1}) &\leq D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\ &= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \\ &= \eta_t \nabla \ell_t(a_t) \cdot (a_t - \tilde{a}_{t+1}) \\ &\leq \eta_t \|\nabla \ell_t(a_t)\|_* \|a_t - \tilde{a}_{t+1}\| \\ &\leq \eta_t \|\nabla \ell_t(a_t)\|_* \|\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})\|_* \end{aligned}$$

where the second equality is from the definition of $\tilde{a}_{t+1}$

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned} D_R(a_t, \tilde{a}_{t+1}) &\le D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\ &= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \\ &= \eta_t \nabla \ell_t(a_t) \cdot (a_t - \tilde{a}_{t+1}) \\ &\le \eta_t \|\nabla \ell_t(a_t)\|_* \|a_t - \tilde{a}_{t+1}\| \\ &\le \eta_t \|\nabla \ell_t(a_t)\|_* \|\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})\|_* \end{aligned}$$

where the second equality is from the definition of $\tilde{a}_{t+1}$
and the second inequality follows from the strong convexity of R wrt $\|\cdot\|$.

Regularization methods: Strongly convex losses

Proof

Also,

$$\begin{aligned}
D_R(a_t, \tilde{a}_{t+1}) &\le D_R(a_t, \tilde{a}_{t+1}) + D_R(\tilde{a}_{t+1}, a_t) \\
&= (\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})) \cdot (a_t - \tilde{a}_{t+1}) \\
&= \eta_t \nabla \ell_t(a_t) \cdot (a_t - \tilde{a}_{t+1}) \\
&\le \eta_t \|\nabla \ell_t(a_t)\|_* \|a_t - \tilde{a}_{t+1}\| \\
&\le \eta_t \|\nabla \ell_t(a_t)\|_* \|\nabla R(a_t) - \nabla R(\tilde{a}_{t+1})\|_* \\
&= \eta_t^2 \|\nabla \ell_t(a_t)\|_*^2,
\end{aligned}$$

where the second equality is from the definition of $\tilde{a}_{t+1}$
and the second inequality follows from the strong convexity of R wrt $\|\cdot\|$.

Regularization methods: Strongly convex losses

Theorem

If ℓ_t is σ-strongly convex wrt R and R is strongly convex wrt $\|\cdot\|$, then for any $a \in \mathcal{A}$, mirror descent, $a_{t+1} = \Pi_{\mathcal{A}}^{R}\left((\nabla R)^{-1}\left(\nabla R(a_t) - \eta_t \nabla \ell_t(a_t)\right)\right)$ with $\eta_t \geq \frac{2}{t\sigma}$ has regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \eta_t \|\nabla \ell_t(a_t)\|_*^2.$$

Regularization methods: Strongly convex losses

Theorem

If ℓ_t is σ-strongly convex wrt R and R is strongly convex wrt $\|\cdot\|$, then for any $a \in \mathcal{A}$, mirror descent, $a_{t+1} = \Pi^R_{\mathcal{A}}\left((\nabla R)^{-1}\left(\nabla R(a_t) - \eta_t \nabla \ell_t(a_t)\right)\right)$ with $\eta_t \geq \frac{2}{t\sigma}$ has regret

$$\sum_{t=1}^n \ell_t(a_t) - \sum_{t=1}^n \ell_t(a) \leq \sum_{t=1}^n \eta_t \|\nabla \ell_t(a_t)\|_*^2.$$

Example

For $R(a) = \frac{1}{2}\|a\|_2^2$, we have

$$\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathbb{R}^d} \sum_{t=1}^n \ell_t(a) \leq \sum_{t=1}^n \eta_t \|\nabla \ell_t\|_*^2 = O\left(\frac{G^2}{\sigma} \log n\right).$$

Brief digression: Linear Losses

Also, even if $\sigma = 0$, this proof shows that we can choose $\eta_t = c/\sqrt{t}$ to get a regret bound of the form

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a))$$
$$\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) + \sum_{t=2}^{n} \left(\frac{\sqrt{t}}{c} - \frac{\sqrt{t-1}}{c} \right) D_R(a, a_t) + \frac{1}{c} D_R(a, a_1)$$

Brief digression: Linear Losses

Also, even if $\sigma = 0$, this proof shows that we can choose $\eta_t = c/\sqrt{t}$ to get a regret bound of the form

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a))$$
$$\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) + \sum_{t=2}^{n} \left(\frac{\sqrt{t}}{c} - \frac{\sqrt{t-1}}{c} \right) D_R(a, a_t) + \frac{1}{c} D_R(a, a_1)$$
$$\leq \sum_{t=1}^{n} \eta_t \|\nabla \ell_t(a_t)\|_*^2 + \frac{D^2}{c} \sum_{t=1}^{n} \left(\sqrt{t} - \sqrt{t-1} \right)$$

Brief digression: Linear Losses

Also, even if $\sigma = 0$, this proof shows that we can choose $\eta_t = c/\sqrt{t}$ to get a regret bound of the form

$$\begin{aligned}
&\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) + \sum_{t=2}^{n} \left(\frac{\sqrt{t}}{c} - \frac{\sqrt{t-1}}{c} \right) D_R(a, a_t) + \frac{1}{c} D_R(a, a_1) \\
&\leq \sum_{t=1}^{n} \eta_t \|\nabla \ell_t(a_t)\|_*^2 + \frac{D^2}{c} \sum_{t=1}^{n} \left(\sqrt{t} - \sqrt{t-1} \right) \\
&\leq \left(cG^2 + \frac{D^2}{c} \right) \sqrt{n}
\end{aligned}$$

Brief digression: Linear Losses

Also, even if $\sigma = 0$, this proof shows that we can choose $\eta_t = c/\sqrt{t}$ to get a regret bound of the form

$$\begin{aligned}
&\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) + \sum_{t=2}^{n} \left(\frac{\sqrt{t}}{c} - \frac{\sqrt{t-1}}{c} \right) D_R(a, a_t) + \frac{1}{c} D_R(a, a_1) \\
&\leq \sum_{t=1}^{n} \eta_t \|\nabla \ell_t(a_t)\|_*^2 + \frac{D^2}{c} \sum_{t=1}^{n} \left(\sqrt{t} - \sqrt{t-1} \right) \\
&\leq \left(cG^2 + \frac{D^2}{c} \right) \sqrt{n} \\
&= O(DG\sqrt{n}).
\end{aligned}$$

Regularization methods: Convexity and Strong Convexity

ℓ_t	η_t	R_n
convex	$\frac{1}{\sqrt{t}}$	$O(\sqrt{n})$
σ-strongly convex	$\frac{1}{\sigma t}$	$O\left(\frac{1}{\sigma}\log n\right)$

Regularization methods: Convexity and Strong Convexity

ℓ_t	η_t	R_n
convex	$\frac{1}{\sqrt{t}}$	$O(\sqrt{n})$
σ-strongly convex	$\frac{1}{\sigma t}$	$O\left(\frac{1}{\sigma}\log n\right)$

All that changes is the step-size.

Regularization methods: Convexity and Strong Convexity

ℓ_t	η_t	R_n
convex	$\frac{1}{\sqrt{t}}$	$O(\sqrt{n})$
σ-strongly convex	$\frac{1}{\sigma t}$	$O\left(\frac{1}{\sigma}\log n\right)$

All that changes is the step-size.
What if we don't know σ?
Can we adapt our step-size to give the right rate?

Outline

1. Binary prediction
2. General online convex
   - Empirical minimization fails
   - Gradient algorithm
   - A regularization viewpoint
   - Bregman divergence
   - Properties of regularization
   - Linearization
   - Mirror descent
   - Regret bounds
   - Strongly convex losses
   - Adaptive regularization
     - Strong convexity (Adaptive Gradient)
     - Diagonal regularizers (AdaGrad)
3. Minimax strategies

Regularization methods: adapting to strong convexity

Adaptive regularization

Regularization methods: adapting to strong convexity

Adaptive regularization

Replace $\ell_t(\cdot)$ with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$,
where g is strongly convex wrt R.

Regularization methods: adapting to strong convexity

Adaptive regularization

Replace $\ell_t(\cdot)$ with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$,
where g is strongly convex wrt R.

$$R_n = \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a))$$

Regularization methods: adapting to strong convexity

Adaptive regularization

Replace $\ell_t(\cdot)$ with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$,
where g is strongly convex wrt R.

$$\begin{aligned} R_n &= \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\ &= \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) + \lambda_t(g(a) - g(a_t))\right) \end{aligned}$$

Regularization methods: adapting to strong convexity

Adaptive regularization

Replace $\ell_t(\cdot)$ with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$,
where g is strongly convex wrt R.

$$
\begin{aligned}
R_n &= \sum_{t=1}^{n} \left(\ell_t(a_t) - \ell_t(a)\right) \\
&= \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) + \lambda_t(g(a) - g(a_t))\right) \\
&\leq D^2 \sum_{t=1}^{n} \lambda_t + \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a)\right),
\end{aligned}
$$

Regularization methods: adapting to strong convexity

Adaptive regularization

Replace $\ell_t(\cdot)$ with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$,
where g is strongly convex wrt R.

$$\begin{aligned} R_n &= \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\ &= \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) + \lambda_t (g(a) - g(a_t)) \right) \\ &\leq D^2 \sum_{t=1}^{n} \lambda_t + \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) \right), \end{aligned}$$

where we've defined $D^2 := \sup_{a,a_t} (g(a) - g(a_t))$.

Regularization methods: adapting to strong convexity

Adaptive regularization

Replace $\ell_t(\cdot)$ with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$,
where g is strongly convex wrt R.

$$
\begin{aligned}
R_n &= \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&= \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) + \lambda_t(g(a) - g(a_t))\right) \\
&\leq D^2 \sum_{t=1}^{n} \lambda_t + \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a)\right),
\end{aligned}
$$

where we've defined $D^2 := \sup_{a,a_t} (g(a) - g(a_t))$.
This is an approximation error term, plus the regret for the regularized losses $\tilde{\ell}_t$.

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

1. How does $\tilde{R}_n$ depend on the λ_ts?

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

1. How does $\tilde{R}_n$ depend on the λ_ts? (We'll give a bound.)

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

1. How does $\tilde{R}_n$ depend on the λ_ts? (We'll give a bound.)
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t?

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

1. How does $\tilde{R}_n$ depend on the λ_ts? (We'll give a bound.)
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t? (Yes.)

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

1. How does $\tilde{R}_n$ depend on the λ_ts? (We'll give a bound.)
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t? (Yes.)
3. Can we choose λ_t online to obtain the best trade-off between these two terms?

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

This is similar to a model selection problem.

1. How does $\tilde{R}_n$ depend on the λ_ts? (We'll give a bound.)
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t? (Yes.)
3. Can we choose λ_t online to obtain the best trade-off between these two terms? (Yes.)

Regularization methods: adapting to strong convexity

Theorem

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

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

and R is strongly convex wrt $\| \cdot \|$,

see, e.g., [B., Hazan, Rakhlin, 2007]

Regularization methods: adapting to strong convexity

Theorem

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

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

and R is strongly convex wrt $\| \cdot \|$, then for any $a \in \mathbb{R}^d$, mirror descent with $\eta_t = 2 / \sum_{s=1}^{t} \sigma_s$ has regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) \leq 2 \sum_{t=1}^{n} \frac{\|\nabla \ell_t(a_t)\|_*^2}{\sum_{s=1}^{t} \sigma_s}.$$

see, e.g., [B., Hazan, Rakhlin, 2007]

Regularization methods: adapting to strong convexity

Theorem

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

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

and R is strongly convex wrt $\|\cdot\|$, then for any $a \in \mathbb{R}^d$, mirror descent with $\eta_t = 2/\sum_{s=1}^{t} \sigma_s$ has regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) \leq 2 \sum_{t=1}^{n} \frac{\|\nabla \ell_t(a_t)\|_*^2}{\sum_{s=1}^{t} \sigma_s}.$$

Notice: η_t is used to update a_t to a_{t+1}, so it uses only past information.

see, e.g., [B., Hazan, Rakhlin, 2007]

Regularization methods: adapting to strong convexity

Proof idea

As before (when σ_t was constant), we have

$$\begin{aligned}
&\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) + \sum_{t=2}^{n} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\sigma_t}{2} \right) D_R(a, a_t) \\
&\quad + \left(\frac{1}{\eta_1} - \frac{\sigma_1}{2} \right) D_R(a, a_1).
\end{aligned}$$

Regularization methods: adapting to strong convexity

Proof idea

As before (when σ_t was constant), we have

$$\begin{aligned}
&\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&\leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, \tilde{a}_{t+1}) + \sum_{t=2}^{n} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\sigma_t}{2} \right) D_R(a, a_t) \\
&\quad + \left(\frac{1}{\eta_1} - \frac{\sigma_1}{2} \right) D_R(a, a_1).
\end{aligned}$$

And the choice of η_t eliminates the second and third terms.

Regularization methods: adapting to strong convexity

Adaptive regularization

Work with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$ (where g is strongly convex wrt R).
If the ℓ_t are σ_t-strongly convex wrt R, then $\tilde{\ell}_t$ are $(\sigma_t + \lambda_t)$-strongly convex.

Regularization methods: adapting to strong convexity

Adaptive regularization

Work with $\tilde{\ell}_t(\cdot) := \ell_t(\cdot) + \lambda_t g(\cdot)$ (where g is strongly convex wrt R). If the ℓ_t are σ_t-strongly convex wrt R, then $\tilde{\ell}_t$ are $(\sigma_t + \lambda_t)$-strongly convex. Using mirror descent for the $\tilde{\ell}_t$s, we choose steps

$$\eta_t = \frac{2}{\sum_{s=1}^{t}(\sigma_s + \lambda_s)}.$$

Regularization methods: adapting to strong convexity

Regret

This strategy incurs regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) \leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) \right)$$

Regularization methods: adapting to strong convexity

Regret

This strategy incurs regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) \leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) \right)$$

$$\leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{\|\nabla \tilde{\ell}_t(a_t)\|_*^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)}$$

Regularization methods: adapting to strong convexity

Regret

This strategy incurs regret

$$
\begin{aligned}
\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) &\leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) \right) \\
&\leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{\|\nabla \tilde{\ell}_t(a_t)\|_*^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)} \\
&\leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)},
\end{aligned}
$$

Regularization methods: adapting to strong convexity

Regret

This strategy incurs regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(a) \le D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \left(\tilde{\ell}_t(a_t) - \tilde{\ell}_t(a) \right)$$
$$\le D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{\|\nabla \tilde{\ell}_t(a_t)\|_*^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)}$$
$$\le D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)},$$

where $\|\nabla \ell_t(a_t)\|_* \le G_t$ and $\|\nabla g(a_t)\|_* \le B$.

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

1. How does $\tilde{R}_n$ depend on the λ_ts?
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t?

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

1. How does $\tilde{R}_n$ depend on the λ_ts?
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t?
3. Can we choose λ_t online to obtain the best trade-off between these two terms?

Regularization methods: adapting to strong convexity

And the best choice of $\lambda_1, \ldots, \lambda_n$ is good here in the convex case:

Example

Assume $\sigma_t \geq 0$. Choose

$$\lambda_1 = \sqrt{\frac{\sum_{t=1}^{n} G_t^2}{B^2 + D^2}}$$

and $\lambda_2 = \cdots = \lambda_n = 0$.

Regularization methods: adapting to strong convexity

And the best choice of $\lambda_1, \ldots, \lambda_n$ is good here in the convex case:

Example

Assume $\sigma_t \geq 0$. Choose

$$\lambda_1 = \sqrt{\frac{\sum_{t=1}^n G_t^2}{B^2 + D^2}}$$

and $\lambda_2 = \cdots = \lambda_n = 0$. Then the bound gives

$$\begin{aligned} R_n &\leq D^2 \sum_{t=1}^n \lambda_t + 2 \sum_{t=1}^n \frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^t (\sigma_s + \lambda_s)} \\ &= O\left(\sqrt{(B^2 + D^2) \sum_{t=1}^n G_t^2} \right). \end{aligned}$$

Regularization methods: adapting to strong convexity

And the best choice of $\lambda_1, \ldots, \lambda_n$ is good here in the convex case:

Example

Assume $\sigma_t \geq 0$. Choose

$$\lambda_1 = \sqrt{\frac{\sum_{t=1}^{n} G_t^2}{B^2 + D^2}}$$

and $\lambda_2 = \cdots = \lambda_n = 0$. Then the bound gives

$$\begin{aligned} R_n &\leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)} \\ &= O\left(\sqrt{(B^2 + D^2) \sum_{t=1}^{n} G_t^2} \right). \end{aligned}$$

If $G_t \leq G$, this is $R_n = O\left(\sqrt{B^2 + D^2}\, G \sqrt{n}\right)$.

Regularization methods: adapting to strong convexity

And the best choice of $\lambda_1, \ldots, \lambda_n$ is good here in the strongly convex case:

Example

Assume $\sigma_t \geq \sigma$ and $G_t \leq G$. Choose $\lambda_1 = \cdots = \lambda_n = 0$.

Regularization methods: adapting to strong convexity

And the best choice of $\lambda_1, \ldots, \lambda_n$ is good here in the strongly convex case:

Example

Assume $\sigma_t \geq \sigma$ and $G_t \leq G$. Choose $\lambda_1 = \cdots = \lambda_n = 0$. Then the bound gives

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + 2 \sum_{t=1}^{n} \frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^{t} (\sigma_s + \lambda_s)}$$
$$= O\left(\frac{G^2}{\sigma} \log n \right).$$

Regularization methods: adapting to strong convexity

We can also obtain a spectrum of rates with the best choice of $\lambda_1, \ldots, \lambda_n$:

Example

Suppose $\sigma_t = t^{-\alpha}$ and $G_t \leq G$. Then the bound gives

$$R_n = \begin{cases} O(\log n) & \text{if } \alpha = 0, \\ O(\sqrt{n}) & \text{if } \alpha > 1/2. \end{cases}$$

Regularization methods: adapting to strong convexity

We can also obtain a spectrum of rates with the best choice of $\lambda_1, \ldots, \lambda_n$:

Example

Suppose $\sigma_t = t^{-\alpha}$ and $G_t \le G$. Then the bound gives

$$R_n = \begin{cases} O(\log n) & \text{if } \alpha = 0, \\ O(n^\alpha) & \text{if } 0 < \alpha \le 1/2, \\ O(\sqrt{n}) & \text{if } \alpha > 1/2. \end{cases}$$

Regularization methods: adapting to strong convexity

We can also obtain a spectrum of rates with the best choice of $\lambda_1, \ldots, \lambda_n$:

Example

Suppose $\sigma_t = t^{-\alpha}$ and $G_t \leq G$. Then the bound gives

$$R_n = \begin{cases} O(\log n) & \text{if } \alpha = 0, \\ O(n^{\alpha}) & \text{if } 0 < \alpha \leq 1/2, \\ O(\sqrt{n}) & \text{if } \alpha > 1/2. \end{cases}$$

(Choose $\lambda_1 = n^{\alpha}$ and $\lambda_2 = \cdots \lambda_n = 0$.)

Regularization methods: adapting to strong convexity

$$R_n \leq D^2 \sum_{t=1}^{n} \lambda_t + \tilde{R}_n(\lambda_1, \ldots, \lambda_n).$$

1. How does $\tilde{R}_n$ depend on the λ_ts?
2. Does the best trade-off between the two terms above ensure the optimal rates for convex and strongly convex ℓ_t?
3. Can we choose λ_t online to obtain the best trade-off between these two terms?

Regularization methods: adapting to strong convexity

Theorem

Choosing

$$\lambda_t = \frac{1}{2}\left(\sqrt{\left(\sum_{s=1}^{t-1}(\sigma_s + \lambda_s) + \sigma_t\right)^2 + \frac{16 G_t^2}{D^2 + B^2}} - \left(\sum_{s=1}^{t-1}(\sigma_s + \lambda_s) + \sigma_t\right)\right)$$

with this regularized mirror descent strategy

[B., Hazan, Rakhlin, 2007]

Regularization methods: adapting to strong convexity

Theorem

Choosing

$$\lambda_t = \frac{1}{2}\left(\sqrt{\left(\sum_{s=1}^{t-1}(\sigma_s + \lambda_s) + \sigma_t\right)^2 + \frac{16G_t^2}{D^2 + B^2}} - \left(\sum_{s=1}^{t-1}(\sigma_s + \lambda_s) + \sigma_t\right)\right)$$

with this regularized mirror descent strategy gives regret

$$R_n = O\left(\inf_{\lambda_1,\ldots,\lambda_n}\left((D^2 + B^2)\sum_{t=1}^{n}\lambda_t + \sum_{t=1}^{n}\frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^{t}(\sigma_s + \lambda_s)}\right)\right).$$

[B., Hazan, Rakhlin, 2007]

Regularization methods: adapting to strong convexity

- Notice that we're using information about each ℓ_t only after we see it.

Regularization methods: adapting to strong convexity

- Notice that we're using information about each ℓ_t only after we see it.
- Compare this to the simple gradient method that we saw earlier, which chooses $\eta = D/(G\sqrt{n})$. Here, we don't need to know the upper bound G (or n): we choose λ_t as a function of information about past losses, and we can compete with the optimal bounds.

Regularization methods: adapting to strong convexity

- Notice that we're using information about each ℓ_t only after we see it.
- Compare this to the simple gradient method that we saw earlier, which chooses $\eta = D/(G\sqrt{n})$. Here, we don't need to know the upper bound G (or n): we choose λ_t as a function of information about past losses, and we can compete with the optimal bounds.
- For instance, for the case of convex functions that satisfy a gradient dual norm bound G,

$$R_n = O\left(\sqrt{B^2 + D^2} G\sqrt{n}\right).$$

(And similarly for the stronger version that replaces G by the rms dual norm of the gradients.)

Regularization methods: adapting to strong convexity

Proof Idea

We prove that balancing the two terms is near-optimal: Consider

$$H_n(\{\lambda_t\}) := \sum_{t=1}^{n} \lambda_t + \sum_{t=1}^{n} \frac{C_t}{\sum_{s=1}^{t}(\sigma_s + \lambda_s)}.$$

Then choosing λ_t to solve the quadratic equation

$$\lambda_t = \frac{C_t}{\sum_{s=1}^{t}(\sigma_s + \lambda_s)}$$

ensures that

$$H_n(\{\lambda_t\}) \leq 2 \inf_{\{\lambda_t^*\}} H_n(\{\lambda_t^*\}).$$

Regularization methods: adapting to strong convexity

Proof Idea

There is an inductive proof of this balancing result, which considers separately the cases

$$\sum_{s=1}^{t} \lambda_s < \sum_{s=1}^{t} \lambda_s^*$$

and

$$\sum_{s=1}^{t} \lambda_s > \sum_{s=1}^{t} \lambda_s^*,$$

and exploits the fact that the two terms of H_t are monotonic in $\sum_{s=1}^{t} \lambda_s$. And the choice of λ_t in the theorem is the positive solution to the appropriate quadratic equation.

Regularization methods: adapting to strong convexity

Theorem

Choosing

$$\lambda_t = \frac{1}{2}\left(\sqrt{\left(\sum_{s=1}^{t-1}(\sigma_s + \lambda_s) + \sigma_t\right)^2 + \frac{16G_t^2}{D^2 + B^2}} - \left(\sum_{s=1}^{t-1}(\sigma_s + \lambda_s) + \sigma_t\right)\right)$$

with this regularized mirror descent strategy gives regret

$$R_n = O\left(\inf_{\lambda_1,\ldots,\lambda_n}\left((D^2 + B^2)\sum_{t=1}^{n}\lambda_t + \sum_{t=1}^{n}\frac{(G_t + \lambda_t B)^2}{\sum_{s=1}^{t}(\sigma_s + \lambda_s)}\right)\right).$$

Outline

1. Binary prediction
2. General online convex
    - Empirical minimization fails
    - Gradient algorithm
    - A regularization viewpoint
    - Bregman divergence
    - Properties of regularization
    - Linearization
    - Mirror descent
    - Regret bounds
    - Strongly convex losses
    - Adaptive regularization
        - Strong convexity (Adaptive Gradient)
        - Diagonal regularizers (AdaGrad)
3. Minimax strategies

Regularization methods: Adaptive regularization

We considered mirror descent where we added an adaptively chosen component of a regularizer g that is strongly convex wrt R. To simplify, assume $g = R$.

Regularization methods: Adaptive regularization

We considered mirror descent where we added an adaptively chosen component of a regularizer g that is strongly convex wrt R. To simplify, assume $g = R$. We can view this in two ways:

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\sum_{s=1}^{t} \eta_s \nabla(\ell_s + \lambda_s R)(a_s) \cdot (a - a_t) + R(a) \right)$$

Regularization methods: Adaptive regularization

We considered mirror descent where we added an adaptively chosen component of a regularizer g that is strongly convex wrt R. To simplify, assume $g = R$. We can view this in two ways:

$$\begin{aligned} a_{t+1} &= \arg\min_{a \in \mathcal{A}} \left(\sum_{s=1}^{t} \eta_s \nabla(\ell_s + \lambda_s R)(a_s) \cdot (a - a_t) + R(a) \right) \\ &= \arg\min_{a \in \mathcal{A}} \left(\eta_t \nabla(\ell_t + \lambda_t R)(a_t) \cdot (a - a_t) + D_R(a, \tilde{a}_t) \right). \end{aligned}$$

Regularization methods: Adaptive regularization

We considered mirror descent where we added an adaptively chosen component of a regularizer g that is strongly convex wrt R. To simplify, assume $g = R$. We can view this in two ways:

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\sum_{s=1}^{t} \eta_s \nabla(\ell_s + \lambda_s R)(a_s) \cdot (a - a_t) + R(a) \right)$$
$$= \arg\min_{a \in \mathcal{A}} \left(\eta_t \nabla(\ell_t + \lambda_t R)(a_t) \cdot (a - a_t) + D_R(a, \tilde{a}_t) \right).$$

Rather than minimizing the sum of the linearization of $\ell_t + \lambda_t R$ plus the regularizer R, we could instead minimize the linearization of ℓ_t plus the regularizer $(1 + \lambda_t)R$:

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta_t \nabla \ell_t(a_t) \cdot (a - a_t) + D_{(1+\lambda_t)R}(a, \tilde{a}_t) \right).$$

Regularization methods: Adaptive regularization

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.
- We could be more ambitious, and consider more than a single parameter (λ_t).

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.
- We could be more ambitious, and consider more than a single parameter (λ_t).
- For example, generalizing the gradient case (where $R(a) = \|a\|_2^2$), we could consider

$$R_t(a) = a^\top M_t a,$$

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.
- We could be more ambitious, and consider more than a single parameter (λ_t).
- For example, generalizing the gradient case (where $R(a) = \|a\|_2^2$), we could consider

$$R_t(a) = a^\top M_t a,$$

  - with $M_t = (1 + \lambda_t)I$ (as before),

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.
- We could be more ambitious, and consider more than a single parameter (λ_t).
- For example, generalizing the gradient case (where $R(a) = \|a\|_2^2$), we could consider

$$R_t(a) = a^\top M_t a,$$

  - with $M_t = (1 + \lambda_t)I$ (as before),
  - with M_t a positive diagonal matrix, or

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.
- We could be more ambitious, and consider more than a single parameter (λ_t).
- For example, generalizing the gradient case (where $R(a) = \|a\|_2^2$), we could consider

$$R_t(a) = a^\top M_t a,$$

  - with $M_t = (1 + \lambda_t)I$ (as before),
  - with M_t a positive diagonal matrix, or
  - with $M_t \succ 0$ (an arbitrary positive definite matrix).

Regularization methods: Adaptive regularization

- Adaptive regularization: $R_t(a) = (1 + \lambda_t)R(a)$.
- We could be more ambitious, and consider more than a single parameter (λ_t).
- For example, generalizing the gradient case (where $R(a) = \|a\|_2^2$), we could consider

$$R_t(a) = a^\top M_t a,$$

  - with $M_t = (1 + \lambda_t)I$ (as before),
  - with M_t a positive diagonal matrix, or
  - with $M_t \succ 0$ (an arbitrary positive definite matrix).
- We can view this as adapting the step-size in different directions.

Regularization methods: Adaptive regularization

Consider the following version of mirror descent (also called 'proximal gradient': stay close to a_t instead of $\tilde{a}_t$):

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

Consider the following version of mirror descent (also called 'proximal gradient': stay close to a_t instead of $\tilde{a}_t$):

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

Similar arguments give the following theorem.

Theorem

For R_t strongly-convex wrt some norm $\| \cdot \|_{R_t}$,

$$R_n \le \frac{1}{\eta} D_{R_1}(a^*, a_1) + \frac{1}{\eta} \sum_{t=1}^{n-1} \left(D_{R_{t+1}}(a, a_{t+1}) - D_{R_t}(a, a_{t+1}) \right) + \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla \ell_t(a_t)\|_{R_t,*}^2 .$$

Example

For $R_t(a) = a^\top M_t a$ with M_t a positive diagonal matrix, say, $M_t = \mathrm{diag}(s_t)$, we have

$$D_{R_t}(a, b) = (a - b)^\top M_t (a - b) = \sum_i (a_i - b_i)^2 s_{t,i}.$$

Example

For $R_t(a) = a^\top M_t a$ with M_t a positive diagonal matrix, say, $M_t = \mathrm{diag}(s_t)$, we have

$$D_{R_t}(a, b) = (a - b)^\top M_t(a - b) = \sum_i (a_i - b_i)^2 s_{t,i}.$$

And D_{R_t} is strongly convex wrt the norm $\|a\|_{R_t}^2 = 2a^\top M_t a$.

Regularization methods: Adaptive regularization

Example

For $R_t(a) = a^\top M_t a$ with M_t a positive diagonal matrix, say, $M_t = \mathrm{diag}(s_t)$, we have

$$D_{R_t}(a, b) = (a - b)^\top M_t (a - b) = \sum_i (a_i - b_i)^2 s_{t,i}.$$

And D_{R_t} is strongly convex wrt the norm $\|a\|^2_{R_t} = 2a^\top M_t a$. Also

$$\|g\|^2_{R_t,*} = \frac{1}{2} g^\top M_t^{-1} g. = \frac{1}{2} \sum_i \frac{g_i^2}{s_{t,i}}.$$

Regularization methods: Adaptive regularization

Example

Applying the theorem, the regret satisfies

$$R_n \leq \frac{1}{\eta} D_{R_1}(a^*, a_1) + \frac{1}{\eta} \sum_{t=1}^{n-1} \left(D_{R_{t+1}}(a, a_{t+1}) - D_{R_t}(a, a_{t+1}) \right)$$

$$+ \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla \ell_t(a_t)\|_{R_t,*}^2$$

Regularization methods: Adaptive regularization

Example

Applying the theorem, the regret satisfies

$$
\begin{aligned}
R_n &\le \frac{1}{\eta} D_{R_1}(a^*, a_1) + \frac{1}{\eta} \sum_{t=1}^{n-1} \left(D_{R_{t+1}}(a, a_{t+1}) - D_{R_t}(a, a_{t+1}) \right) \\
&\quad + \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla \ell_t(a_t)\|_{R_t,*}^2 \\
&\le \frac{1}{\eta} D_{R_1}(a^*, a_1) + \frac{1}{\eta} \sum_{t=1}^{n-1} \max_i (a_i^* - a_{t+1,i})^2 \|s_{t+1} - s_t\|_1 \\
&\quad + \frac{\eta}{4} \sum_{t=1}^{n} \nabla \ell_t(a_t)^\top \, \mathrm{diag}(s_t)^{-1} \nabla \ell_t(a_t).
\end{aligned}
$$

Regularization methods: Adaptive regularization

Adagrad

[Duchi, Hazan, Singer, 2011]

If we insist that the regularization increases (that is, the components of s_t are monotonically non-decreasing with t), we can choose

$$s_{t,i} = \sqrt{\sum_{s=1}^{t} \nabla \ell_t(a_t)_i^2},$$

Regularization methods: Adaptive regularization

Adagrad

[Duchi, Hazan, Singer, 2011]

If we insist that the regularization increases (that is, the components of s_t are monotonically non-decreasing with t), we can choose

$$s_{t,i} = \sqrt{\sum_{s=1}^{t} \nabla \ell_t(a_t)_i^2},$$

$$\eta = D_\infty := \sup_{a^*, a_t} \|a^* - a_t\|_\infty,$$

Regularization methods: Adaptive regularization

Adagrad

[Duchi, Hazan, Singer, 2011]

If we insist that the regularization increases (that is, the components of s_t are monotonically non-decreasing with t), we can choose

$$s_{t,i} = \sqrt{\sum_{s=1}^{t} \nabla \ell_t(a_t)_i^2},$$

$$\eta = D_\infty := \sup_{a^*, a_t} \|a^* - a_t\|_\infty,$$

to give an adaptivity result (versus *constant* s):

$$R_n \leq c \min_{\eta, s} \left(\frac{D_\infty^2}{\eta} s^\top \mathbf{1} + \eta \sum_{t=1}^{n} \nabla \ell_t(a_t)^\top \operatorname{diag}(s)^{-1} \nabla \ell_t(a_t) \right)$$

Regularization methods: Adaptive regularization

Adagrad

[Duchi, Hazan, Singer, 2011]

If we insist that the regularization increases (that is, the components of s_t are monotonically non-decreasing with t), we can choose

$$s_{t,i} = \sqrt{\sum_{s=1}^{t} \nabla \ell_t(a_t)_i^2},$$

$$\eta = D_\infty := \sup_{a^*, a_t} \|a^* - a_t\|_\infty,$$

to give an adaptivity result (versus *constant* s):

$$R_n \leq c \min_{\eta, s} \left(\frac{D_\infty^2}{\eta} s^\top \mathbf{1} + \eta \sum_{t=1}^{n} \nabla \ell_t(a_t)^\top \operatorname{diag}(s)^{-1} \nabla \ell_t(a_t) \right)$$

$$= O\left(D_\infty \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{n} \nabla \ell_t(a_t)_i^2} \right).$$

Regularization methods: Adaptive regularization

Adagrad

Regularization methods: Adaptive regularization

Adagrad

- The gradient term might be much smaller than $\sqrt{nd}$.

Adagrad

- The gradient term might be much smaller than $\sqrt{nd}$.
- For instance, if the gradients are sparse and bounded (for instance, for logistic regression with sparse $\{0, 1\}$-valued features), then we expect the gradient terms to be much smaller.
  For features that appear more frequently, the $s_{t,i}$ will be larger (learning rate slower in those directions).

Regularization methods: Adaptive regularization

Adagrad

- The gradient term might be much smaller than $\sqrt{nd}$.
- For instance, if the gradients are sparse and bounded (for instance, for logistic regression with sparse $\{0, 1\}$-valued features), then we expect the gradient terms to be much smaller.
  For features that appear more frequently, the $s_{t,i}$ will be larger (learning rate slower in those directions).
- More generally, for coordinate directions with large gradients, we can make the corresponding component of s large (to keep things more stable in those directions), and for coordinate directions with small gradients, we can use less regularization.

Regularization methods: Adaptive regularization

Adagrad

A similar approach can be applied to matrices, with

$$M_t = \frac{\left(\sum_{s=1}^{t} \nabla \ell_t(a_t) \nabla \ell_t(a_t)^\top\right)^{1/2}}{\mathrm{tr}\left(\sum_{s=1}^{t} \nabla \ell_t(a_t) \nabla \ell_t(a_t)^\top\right)^{1/2}}$$

playing the role of s_t.

Outline

1. Binary prediction
2. General online convex
    - Empirical minimization fails
    - Gradient algorithm
    - A regularization viewpoint
    - Bregman divergence
    - Properties of regularization
    - Linearization
    - Mirror descent
    - Regret bounds
    - Strongly convex losses
    - Adaptive regularization
3. Minimax strategies

Outline

1. Binary prediction
2. General online convex
3. Minimax strategies
   - Convex and strongly convex losses
   - The linear game

Convex and strongly convex losses

The convex and linear games

For a convex set $\mathcal{A} \subset \mathbb{R}^d$ and a sequence $G_1, \ldots, G_n \geq 0$, define $\mathcal{G}_{conv}(\mathcal{A}, \{G_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \{\ell : \|\nabla \ell(a_t)\| \leq G_t, \ \ell \text{ convex}\}.$$

Convex and strongly convex losses

The convex and linear games

For a convex set $\mathcal{A} \subset \mathbb{R}^d$ and a sequence $G_1, \ldots, G_n \geq 0$, define $\mathcal{G}_{conv}(\mathcal{A}, \{G_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \{\ell : \|\nabla \ell(a_t)\| \leq G_t, \;\; \ell \text{ convex}\}.$$

Define $\mathcal{G}_{lin}(\mathcal{A}, \{G_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \ell(a) = v^\top (a - a_t) + c,\, v \in \mathbb{R}^d,\, c \in \mathbb{R},\, \|v\| \leq G_t\right\}.$$

Convex and strongly convex losses

The convex and linear games

For a convex set $\mathcal{A} \subset \mathbb{R}^d$ and a sequence $G_1, \ldots, G_n \geq 0$, define $\mathcal{G}_{conv}(\mathcal{A}, \{G_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \{\ell : \|\nabla \ell(a_t)\| \leq G_t,\ \ \ell \text{ convex}\}.$$

Define $\mathcal{G}_{lin}(\mathcal{A}, \{G_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \ell(a) = v^\top (a - a_t) + c,\, v \in \mathbb{R}^d,\, c \in \mathbb{R},\, \|v\| \leq G_t \right\}.$$

- The adversary's constraints depend on the player's choices.

Convex and strongly convex losses

The strongly convex and quadratic games

For a convex set $\mathcal{A} \subset \mathbb{R}^d$ and sequences $G_1, \ldots, G_n \geq 0$ and $\sigma_1, \ldots, \sigma_n \geq 0$, define $\mathcal{G}_{st-conv}\left(\mathcal{A}, \{G_t\}, \{\sigma_t\}\right)$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \|\nabla \ell(a_t)\| \leq G_t,\ \nabla^2 \ell \succeq \sigma_t I\right\}.$$

Convex and strongly convex losses

The strongly convex and quadratic games

For a convex set $\mathcal{A} \subset \mathbb{R}^d$ and sequences $G_1, \ldots, G_n \geq 0$ and $\sigma_1, \ldots, \sigma_n \geq 0$, define $\mathcal{G}_{st-conv}(\mathcal{A}, \{G_t\}, \{\sigma_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \|\nabla \ell(a_t)\| \leq G_t,\ \nabla^2 \ell \succeq \sigma_t I\right\}.$$

Define $\mathcal{G}_{quad}(\mathcal{A}, \{G_t\}, \{\sigma_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2}\|a - a_t\|^2 + c,\ v \in \mathbb{R}^d,\ c \in \mathbb{R},\ \|v\| \leq G_t\right\}.$$

Convex and strongly convex losses

The strongly convex and quadratic games

For a convex set $\mathcal{A} \subset \mathbb{R}^d$ and sequences $G_1, \ldots, G_n \geq 0$ and $\sigma_1, \ldots, \sigma_n \geq 0$, define $\mathcal{G}_{st-conv}(\mathcal{A}, \{G_t\}, \{\sigma_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \|\nabla \ell(a_t)\| \leq G_t,\ \nabla^2 \ell \succeq \sigma_t I\right\}.$$

Define $\mathcal{G}_{quad}(\mathcal{A}, \{G_t\}, \{\sigma_t\})$ as the online convex optimization game with constraints $a_t \in \mathcal{A}$ and

$$\ell_t \in \left\{\ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2}\|a - a_t\|^2 + c,\ v \in \mathbb{R}^d,\ c \in \mathbb{R},\ \|v\| \leq G_t\right\}.$$

- Again, the adversary's constraints depend on the player's choices.

Convex and strongly convex losses

Theorem

For fixed $\mathcal{A}$, $\{G_t\}$ and $\{\sigma_t\}$, we have

$$V_n\left(\mathcal{G}_{st-conv}\left(\mathcal{A},\{G_t\},\{\sigma_t\}\right)\right) = V_n\left(\mathcal{G}_{quad}\left(\mathcal{A},\{G_t\},\{\sigma_t\}\right)\right),$$

[Abernethy, B., Rakhlin, Tewari, 2008]

Convex and strongly convex losses

Theorem

For fixed $\mathcal{A}$, $\{G_t\}$ and $\{\sigma_t\}$, we have

$$V_n(\mathcal{G}_{st-conv}(\mathcal{A}, \{G_t\}, \{\sigma_t\})) = V_n(\mathcal{G}_{quad}(\mathcal{A}, \{G_t\}, \{\sigma_t\})),$$

$$V_n(\mathcal{G}_{conv}(\mathcal{A}, \{G_t\})) = V_n(\mathcal{G}_{lin}(\mathcal{A}, \{G_t\})).$$

[Abernethy, B., Rakhlin, Tewari, 2008]

Convex and strongly convex losses

Lemma

Fix sets $N_1, \ldots, N_n$ and $M \subseteq N_t$.
Suppose that for all $\ell_t \in N_t$ and $a_t \in \mathcal{A}$ there is an $\ell_t^* \in M$ such that for all $a_1, \ell_1, \ldots, a_{t-1}, \ell_{t-1}$, and $a_{t+1}, \ell_{t+1}, \ldots, a_n, \ell_n$,

$$R_n(a_1, \ell_1, \ldots, a_t, \ell_t, \ldots, a_n, \ell_n) \leq R_n(a_1, \ell_1, \ldots, a_t, \ell_t^*, \ldots, a_n, \ell_n).$$

Convex and strongly convex losses

Lemma

Fix sets $N_1, \ldots, N_n$ and $M \subseteq N_t$.
Suppose that for all $\ell_t \in N_t$ and $a_t \in \mathcal{A}$ there is an $\ell_t^* \in M$ such that for all $a_1, \ell_1, \ldots, a_{t-1}, \ell_{t-1}$, and $a_{t+1}, \ell_{t+1}, \ldots, a_n, \ell_n$,

$$R_n(a_1, \ell_1, \ldots, a_t, \ell_t, \ldots, a_n, \ell_n) \leq R_n(a_1, \ell_1, \ldots, a_t, \ell_t^*, \ldots, a_n, \ell_n).$$

Then

$$\inf_{a_1 \in \mathcal{A}} \sup_{\ell_1 \in N_1} \cdots \inf_{a_t \in \mathcal{A}} \sup_{\ell_t \in N_t} \cdots \inf_{a_n \in \mathcal{A}} \sup_{\ell_n \in N_n} R_n(a_1, \ell_1, \ldots, a_n, \ell_n)$$

Convex and strongly convex losses

Lemma

Fix sets $N_1, \ldots, N_n$ and $M \subseteq N_t$.
Suppose that for all $\ell_t \in N_t$ and $a_t \in \mathcal{A}$ there is an $\ell_t^* \in M$ such that for all $a_1, \ell_1, \ldots, a_{t-1}, \ell_{t-1}$, and $a_{t+1}, \ell_{t+1}, \ldots, a_n, \ell_n$,

$$R_n(a_1, \ell_1, \ldots, a_t, \ell_t, \ldots, a_n, \ell_n) \leq R_n(a_1, \ell_1, \ldots, a_t, \ell_t^*, \ldots, a_n, \ell_n).$$

Then

$$\begin{aligned}
&\inf_{a_1 \in \mathcal{A}} \sup_{\ell_1 \in N_1} \cdots \inf_{a_t \in \mathcal{A}} \sup_{\ell_t \in N_t} \cdots \inf_{a_n \in \mathcal{A}} \sup_{\ell_n \in N_n} R_n(a_1, \ell_1, \ldots, a_n, \ell_n) \\
&= \inf_{a_1 \in \mathcal{A}} \sup_{\ell_1 \in N_1} \cdots \inf_{a_t \in \mathcal{A}} \sup_{\ell_t \in M} \cdots \inf_{a_n \in \mathcal{A}} \sup_{\ell_n \in N_n} R_n(a_1, \ell_1, \ldots, a_n, \ell_n).
\end{aligned}$$

Convex and strongly convex losses

Lemma

Fix sets $N_1, \ldots, N_n$ and $M \subseteq N_t$.
Suppose that for all $\ell_t \in N_t$ and $a_t \in \mathcal{A}$ there is an $\ell_t^* \in M$ such that for all $a_1, \ell_1, \ldots, a_{t-1}, \ell_{t-1}$, and $a_{t+1}, \ell_{t+1}, \ldots, a_n, \ell_n$,

$$R_n(a_1, \ell_1, \ldots, a_t, \ell_t, \ldots, a_n, \ell_n) \leq R_n(a_1, \ell_1, \ldots, a_t, \ell_t^*, \ldots, a_n, \ell_n).$$

Then

$$\begin{aligned} &\inf_{a_1 \in \mathcal{A}} \sup_{\ell_1 \in N_1} \cdots \inf_{a_t \in \mathcal{A}} \sup_{\ell_t \in N_t} \cdots \inf_{a_n \in \mathcal{A}} \sup_{\ell_n \in N_n} R_n(a_1, \ell_1, \ldots, a_n, \ell_n) \\ &= \inf_{a_1 \in \mathcal{A}} \sup_{\ell_1 \in N_1} \cdots \inf_{a_t \in \mathcal{A}} \sup_{\ell_t \in M} \cdots \inf_{a_n \in \mathcal{A}} \sup_{\ell_n \in N_n} R_n(a_1, \ell_1, \ldots, a_n, \ell_n). \end{aligned}$$

(Because $M \subset N_t$, and it contains ℓ_t^* that's always at least as good as ℓ_t.)

Convex and strongly convex losses

Proof idea

For the strongly convex case, define

$$M := \left\{ \ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2 + c,\ \|v\| \leq G_t \right\},$$

Convex and strongly convex losses

Proof idea

For the strongly convex case, define

$$M := \left\{ \ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2 + c,\ \|v\| \le G_t \right\},$$

and notice that

$$M \subseteq N_t := \left\{ \ell : \|\nabla \ell(a_t)\| \le G_t,\ \nabla^2 \ell \succeq \sigma_t I \right\}.$$

Convex and strongly convex losses

Proof idea

For the strongly convex case, define

$$M := \left\{ \ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2 + c,\ \|v\| \le G_t \right\},$$

and notice that

$$M \subseteq N_t := \left\{ \ell : \|\nabla \ell(a_t)\| \le G_t,\ \nabla^2 \ell \succeq \sigma_t I \right\}.$$

For $\ell_t \in N_t$, define ℓ_t^* as

$$\ell_t^*(a) = \ell_t(a_t) + \nabla \ell_t(a_t)^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2.$$

Convex and strongly convex losses

Proof idea

For the strongly convex case, define

$$M := \left\{ \ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2 + c,\ \|v\| \leq G_t \right\},$$

and notice that

$$M \subseteq N_t := \left\{ \ell : \|\nabla \ell(a_t)\| \leq G_t,\ \nabla^2 \ell \succeq \sigma_t I \right\}.$$

For $\ell_t \in N_t$, define ℓ_t^* as

$$\ell_t^*(a) = \ell_t(a_t) + \nabla \ell_t(a_t)^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2.$$

Notice that $\ell_t^* \in M$, since $\ell_t^*(a_t) = \ell_t(a_t)$ and $\nabla \ell_t(a_t) = \nabla \ell_t^*(a_t)$.

Convex and strongly convex losses

Proof idea

For the strongly convex case, define

$$M := \left\{ \ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2 + c,\ \|v\| \le G_t \right\},$$

and notice that

$$M \subseteq N_t := \left\{ \ell : \|\nabla \ell(a_t)\| \le G_t,\ \nabla^2 \ell \succeq \sigma_t I \right\}.$$

For $\ell_t \in N_t$, define ℓ_t^* as

$$\ell_t^*(a) = \ell_t(a_t) + \nabla \ell_t(a_t)^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2.$$

Notice that $\ell_t^* \in M$, since $\ell_t^*(a_t) = \ell_t(a_t)$ and $\nabla \ell_t(a_t) = \nabla \ell_t^*(a_t)$. Also, $\ell_t(a) \ge \ell_t^*(a)$ for all a, so M and N_t satisfy the conditions of the lemma.

Convex and strongly convex losses

Proof idea

For the strongly convex case, define

$$M := \left\{ \ell : \ell(a) = v^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2 + c,\ \|v\| \le G_t \right\},$$

and notice that

$$M \subseteq N_t := \left\{ \ell : \|\nabla \ell(a_t)\| \le G_t,\ \nabla^2 \ell \succeq \sigma_t I \right\}.$$

For $\ell_t \in N_t$, define ℓ_t^* as

$$\ell_t^*(a) = \ell_t(a_t) + \nabla \ell_t(a_t)^\top (a - a_t) + \frac{\sigma_t}{2} \|a - a_t\|^2.$$

Notice that $\ell_t^* \in M$, since $\ell_t^*(a_t) = \ell_t(a_t)$ and $\nabla \ell_t(a_t) = \nabla \ell_t^*(a_t)$. Also, $\ell_t(a) \ge \ell_t^*(a)$ for all a, so M and N_t satisfy the conditions of the lemma. The convex/linear case uses a similar argument.

Outline

1. Binary prediction
2. General online convex
3. Minimax strategies
    - Convex and strongly convex losses
    - The linear game

The linear game

Theorem

For $\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq r\}$ with $d \geq 3$, and a fixed sequence $\{G_t\}$,

$$V_n\left(\mathcal{G}_{conv}\left(\mathcal{A}, \{G_t\}\right)\right) = V_n\left(\mathcal{G}_{lin}\left(\mathcal{A}, \{G_t\}\right)\right) = r\sqrt{\sum_{t=1}^{n} G_t^2}.$$

[Abernethy, B., Rakhlin, Tewari, 2008]

The linear game

Proof

1. Wlog, we can assume $r = 1$ and $\ell_t(a) = w^\top a$ with $\|w\| \leq G_t$.

The linear game

Proof

1. Wlog, we can assume $r = 1$ and $\ell_t(a) = w^\top a$ with $\|w\| \leq G_t$.
2. Writing $W_t := \sum_{s=1}^{t} w_s$,

$$\min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) = -\|W_n\|.$$

The linear game

Proof

3. The adversary can ensure

$$R_n \geq \sqrt{\sum_{t=1}^{n} G_t^2},$$

by playing w_t satisfying

$$w_t^\top a_t = 0, \quad w_t^\top W_{t-1} = 0, \quad \|w_t\| = G_t.$$

The linear game

Proof

4. The adversary can ensure

$$R_n \geq \sqrt{\sum_{t=1}^{n} G_t^2},$$

by playing w_t satisfying

$$w_t^\top a_t = 0, \quad w_t^\top W_{t-1} = 0, \quad \|w_t\| = G_t.$$

To see this, notice that this choice ensures $\sum_{t=1}^{n} \ell_t(a_t) = 0$ and so $R_n = \|W_n\|$.

The linear game

Proof

5. The adversary can ensure

$$R_n \geq \sqrt{\sum_{t=1}^{n} G_t^2},$$

by playing w_t satisfying

$$w_t^\top a_t = 0, \quad w_t^\top W_{t-1} = 0, \quad \|w_t\| = G_t.$$

To see this, notice that this choice ensures $\sum_{t=1}^{n} \ell_t(a_t) = 0$ and so $R_n = \|W_n\|$. But

$$\|W_t\| = \|W_{t-1} + w_t\| = \sqrt{\|W_{t-1}\|^2 + \|w_t\|^2} = \sqrt{\sum_{s=1}^{t} G_s^2}.$$

The linear game

Proof

6. If the player defines $W_0 = 0$ and chooses

$$a_t = \frac{-W_{t-1}}{\sqrt{\|W_{t-1}\|^2 + \sum_{s=t}^{n} G_s^2}},$$

then

$$R_n \leq \sqrt{\sum_{t=1}^{n} G_t^2}.$$

The linear game

Proof

This is equivalent to showing that, for this a_t, no matter what choices of w_t the adversary makes,

$$\sum_{t=1}^{n} w_t^\top a_t + \|W_n\| \le \sqrt{\sum_{t=1}^{n} G_t^2}.$$

The linear game

Proof

This is equivalent to showing that, for this a_t, no matter what choices of w_t the adversary makes,

$$\sum_{t=1}^{n} w_t^\top a_t + \|W_n\| \leq \sqrt{\sum_{t=1}^{n} G_t^2}.$$

The proof is by a backward induction, and involves a 2-dimensional geometric argument (since a_t is aligned with W_{t-1}, we need only consider the role of w_t).

Outline

1. Binary prediction
2. General online convex
3. Minimax strategies
   - Convex and strongly convex losses
   - The linear game