Regret minimization via online mirror descent

Ahmed El Alaoui
Electrical Engineering and Computer Science
UC Berkeley

October 7, 2013

The experts framework

A sequential decision problem:

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

1. At round t, each expert provides a suggestion $i \in \{1, \cdots, k\}$.

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

1. At round t, each expert provides a suggestion $i \in \{1, \cdots, k\}$.
2. The forecaster follows one expert. If decision i is made, incur a loss $l_{t,i}$.

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

1. At round t, each expert provides a suggestion $i \in \{1, \cdots, k\}$.
2. The forecaster follows one expert. If decision i is made, incur a loss $l_{t,i}$.
3. The full vector of losses $l_t = (l_{t,i})_{1 \leq i \leq k}$ is then revealed.

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

1. At round t, each expert provides a suggestion $i \in \{1, \cdots, k\}$.
2. The forecaster follows one expert. If decision i is made, incur a loss $l_{t,i}$.
3. The full vector of losses $l_t = (l_{t,i})_{1 \leq i \leq k}$ is then revealed.
4. Repeat.

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

1. At round t, each expert provides a suggestion $i \in \{1, \cdots, k\}$.
2. The forecaster follows one expert. If decision i is made, incur a loss $l_{t,i}$.
3. The full vector of losses $l_t = (l_{t,i})_{1 \leq i \leq k}$ is then revealed.
4. Repeat.

After T steps, define the regret of any strategy $\mathcal{A}$ relative the best (stationary) strategy

$$R(\mathcal{A}) = \sum_{t=1}^{T} l_{t,i_t} - \min_{1 \leq i \leq k} \sum_{t=1}^{T} l_{t,i}.$$

The experts framework

A sequential decision problem:The forecaster has a set of k decisions associated with a set of *experts* $\{1, \cdots, k\}$.

1. At round t, each expert provides a suggestion $i \in \{1, \cdots, k\}$.
2. The forecaster follows one expert. If decision i is made, incur a loss $l_{t,i}$.
3. The full vector of losses $l_t = (l_{t,i})_{1 \leq i \leq k}$ is then revealed.
4. Repeat.

After T steps, define the regret of any strategy $\mathcal{A}$ relative the best (stationary) strategy

$$R(\mathcal{A}) = \sum_{t=1}^{T} l_{t,i_t} - \min_{1 \leq i \leq k} \sum_{t=1}^{T} l_{t,i}.$$

Objective:

$$\text{minimize}_{\mathcal{A}=\{i_1, \cdots, i_T\}} R(\mathcal{A}).$$

The experts framework

Problem: For any $\mathcal{A} = \{i_1, \cdots, i_T\}$ are picked deterministically, there exist an instance such that $R(\mathcal{A}) = \Omega(T)$.

The experts framework

Problem: For any $\mathcal{A} = \{i_1, \cdots, i_T\}$ are picked deterministically, there exist an instance such that $R(\mathcal{A}) = \Omega(T)$.

Solution: Randomize!

The experts framework

Problem: For any $\mathcal{A} = \{i_1, \cdots, i_T\}$ are picked deterministically, there exist an instance such that $R(\mathcal{A}) = \Omega(T)$.

Solution: Randomize!

Multiplicative Weights Update (MWU) Algorithm:

1. At time $t = 0$, set $w_0 = \frac{1}{n}(1, \cdots, 1)^T$.
2. At time $t \geq 1$, choose step size γ_t, $\forall i \in \{1, \cdots, k\}$, set

$$w_{t+1,i} = \frac{1}{Z_t} w_{t,i} \exp(-\gamma_t l_{t,i})$$

where

$$Z_t = \sum_{i=1}^{k} w_{t,i} \exp(-\gamma_t l_{t,i}).$$

The experts framework

Bound on (MWU):

Theorem 1 (Kakade).

If the losses are bounded and $\gamma_t = \gamma = 1/\sqrt{T \log k}$*, the expected regret is bounded by* $O(\sqrt{T})$*:*

$$\mathbb{E}[R(\mathcal{A})] = \sum_{t=1}^{T} \mathbb{E}_{i_t}[l_{t,i_t}] - \min_{1 \le i \le k} \sum_{t=1}^{T} l_{t,i} \le 2\sqrt{T \log k}.$$

The experts framework

Bound on (MWU):

Theorem 1 (Kakade).

If the losses are bounded and $\gamma_t = \gamma = 1/\sqrt{T \log k}$*, the expected regret is bounded by* $O(\sqrt{T})$*:*

$$\mathbb{E}[R(\mathcal{A})] = \sum_{t=1}^{T} \mathbb{E}_{i_t}[l_{t,i_t}] - \min_{1 \le i \le k} \sum_{t=1}^{T} l_{t,i} \le 2\sqrt{T \log k}.$$

This rate is optimal.

Formulating a stochastic objective

Now we focus on the objective we seek to minimize:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i_t}[l_{t,i_t}]$$

Formulating a stochastic objective

Now we focus on the objective we seek to minimize:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i_t}[l_{t,i_t}] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} w_{t,i}\, l_{t,i}$$

Formulating a stochastic objective

Now we focus on the objective we seek to minimize:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i_t}[l_{t,i_t}] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} w_{t,i}\, l_{t,i}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\omega_t \cdot l_t$$

Formulating a stochastic objective

Now we focus on the objective we seek to minimize:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i_t}[l_{t,i_t}] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} w_{t,i}\, l_{t,i}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\omega_t \cdot l_t = \frac{1}{T}\sum_{t=1}^{T} F(\omega_t, l_t).$$

Where $F : \mathcal{X} \times \Xi \to \mathbb{R}, (x, \xi) \to x \cdot \xi$. (Here $x := \omega$ and $\xi := l$)

Formulating a stochastic objective

Now we focus on the objective we seek to minimize:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i_t}[l_{t,i_t}] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} w_{t,i}\, l_{t,i}$$
$$= \frac{1}{T}\sum_{t=1}^{T}\omega_t \cdot l_t = \frac{1}{T}\sum_{t=1}^{T} F(\omega_t, l_t).$$

Where $F : \mathcal{X} \times \Xi \to \mathbb{R}, (x, \xi) \to x \cdot \xi$. (Here $x := \omega$ and $\xi := l$)
Now consider the problem

$$\text{minimize}_{x \in X}\, \hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} F(x, \xi_t)$$

A detour by stochastic optimization

Recall the general from of a stochastic optimisation problem:

$$\text{minimize}_{x \in X} f(x) = \mathbb{E}_{\xi}[F(x, \xi)]$$

where

- the decision variable is $x \in X$ (convex closed and bounded)

A detour by stochastic optimization

Recall the general from of a stochastic optimisation problem:

$$\text{minimize}_{x \in X} f(x) = \mathbb{E}_{\xi}[F(x, \xi)]$$

where

- the decision variable is $x \in X$ (convex closed and bounded)
- ξ is a random vector of **unknown** and well-behaved distribution P.

A detour by stochastic optimization

Recall the general from of a stochastic optimisation problem:

$$\text{minimize}_{x \in X} f(x) = \mathbb{E}_{\xi}[F(x, \xi)]$$

where

- the decision variable is $x \in X$ (convex closed and bounded)
- ξ is a random vector of **unknown** and well-behaved distribution P.
- Nevertheless, we can access partial information about P by sampling from it.

A detour by stochastic optimization

Let $\xi_{[T]} = (\xi_1, \cdots, \xi_T)$ be a set of iid samples from P.

A detour by stochastic optimization

Let $\xi_{[T]} = (\xi_1, \cdots, \xi_T)$ be a set of iid samples from P.

Approach 1: Approximate the true problem by the sample average problem

$$\text{minimize}_{x \in X} \hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} F(x, \xi_t).$$

A detour by stochastic optimization

Let $\xi_{[T]} = (\xi_1, \cdots, \xi_T)$ be a set of iid samples from P.

Approach 1: Approximate the true problem by the sample average problem

$$\text{minimize}_{x \in X} \hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} F(x, \xi_t).$$

Then solve this problem using a numerical procedure of your choice.

A detour by stochastic optimization

Let $\xi_{[T]} = (\xi_1, \cdots, \xi_T)$ be a set of iid samples from P.

Approach 1: Approximate the true problem by the sample average problem

$$\text{minimize}_{x \in X} \hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} F(x, \xi_t).$$

Then solve this problem using a numerical procedure of your choice.

Caveats:

- Need a large nb of samples *before* starting to optimize.

A detour by stochastic optimization

Let $\xi_{[T]} = (\xi_1, \cdots, \xi_T)$ be a set of iid samples from P.

Approach 1: Approximate the true problem by the sample average problem

$$\text{minimize}_{x \in X} \hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} F(x, \xi_t).$$

Then solve this problem using a numerical procedure of your choice.

Caveats:

- Need a large nb of samples *before* starting to optimize.
- Can only reduce approximation error (statistical error remains untouched).

A detour by stochastic optimization

Let $\xi_{[T]} = (\xi_1, \cdots, \xi_T)$ be a set of iid samples from P.

Approach 1: Approximate the true problem by the sample average problem

$$\text{minimize}_{x \in X} \hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} F(x, \xi_t).$$

Then solve this problem using a numerical procedure of your choice.

Caveats:

- Need a large nb of samples *before* starting to optimize.
- Can only reduce approximation error (statistical error remains untouched).
- Unsuitable in a streaming context.

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally.

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally. Here enter the ideas of stochastic gradient descent.

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally. Here enter the ideas of stochastic gradient descent.

- Suppose there is a mechanism (oracle) which for a given input x returns an unbiased estimate $G(x, \xi)$ of $\nabla f(x)$ (i.e. $\nabla f(x) = \mathbb{E}[G(x, \xi)]$).

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally. Here enter the ideas of stochastic gradient descent.

- Suppose there is a mechanism (oracle) which for a given input x returns an unbiased estimate $G(x, \xi)$ of $\nabla f(x)$ (i.e. $\nabla f(x) = \mathbb{E}[G(x, \xi)]$).
- Consider the iteration

$$x_{t+1} = \Pi_X(x_t - \gamma_t G(x, \xi_t)) \qquad \text{(SGD)}$$

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally. Here enter the ideas of stochastic gradient descent.

- Suppose there is a mechanism (oracle) which for a given input x returns an unbiased estimate $G(x, \xi)$ of $\nabla f(x)$ (i.e. $\nabla f(x) = \mathbb{E}[G(x, \xi)]$).
- Consider the iteration

$$x_{t+1} = \Pi_X(x_t - \gamma_t G(x, \xi_t)) \qquad \text{(SGD)}$$

Theorem 2 (Nemirovski, Yudin).

If f convex, X convex closed and bounded then

$$\mathbb{E}[f(x_t) - f(x^*)] \leq O(1/\sqrt{t}).$$

$O(1/t)$ if f is strongly convex with Lipschitz gradient.

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally. Here enter the ideas of stochastic gradient descent.

- Suppose there is a mechanism (oracle) which for a given input x returns an unbiased estimate $G(x, \xi)$ of $\nabla f(x)$ (i.e. $\nabla f(x) = \mathbb{E}[G(x, \xi)]$).
- Consider the iteration

$$x_{t+1} = \Pi_X(x_t - \gamma_t G(x, \xi_t)) \qquad \text{(SGD)}$$

Theorem 2 (Nemirovski, Yudin).

If f convex, X convex closed and bounded then

$$\mathbb{E}[f(x_t) - f(x^*)] \leq O(1/\sqrt{t}).$$

$O(1/t)$ if f is strongly convex with Lipschitz gradient.

The right rate of convergence ...

A detour by stochastic optimization

Approach 2: Directly tackle original problem by leveraging the information carried by each sample received and make a decision incrementally. Here enter the ideas of stochastic gradient descent.

- Suppose there is a mechanism (oracle) which for a given input x returns an unbiased estimate $G(x, \xi)$ of $\nabla f(x)$ (i.e. $\nabla f(x) = \mathbb{E}[G(x, \xi)]$).
- Consider the iteration

$$x_{t+1} = \Pi_X(x_t - \gamma_t G(x, \xi_t)) \qquad \text{(SGD)}$$

Theorem 2 (Nemirovski, Yudin).

If f convex, X convex closed and bounded then

$$\mathbb{E}[f(x_t) - f(x^*)] \leq O(1/\sqrt{t}).$$

$O(1/t)$ *if f is strongly convex with Lipschitz gradient.*

The right rate of convergence ... but the wrong iteration scheme!

Stochastic Mirror Descent

Need to generalize the notion of gradient descent.

Stochastic Mirror Descent

Need to generalize the notion of gradient descent. Ingredients:

- A general norm $\|\cdot\|$ on $\mathbb{R}^n$
- A *distance-generating function* $\omega : X \to \mathbb{R}$ continuous and strongly convex
- Define the *prox-function* (or Bregman divergence) $V : X \times X \to \mathbb{R}$ as

$$V(x, z) = \omega(z) - \omega(x) - \nabla\omega(x)^T(z - x).$$

- For $x \in X$ define the *prox-mapping* $P_x : \mathbb{R}^n \to X$ as

$$P_x(y) = \arg\min_{x \in X}\{y^T(z - x) + V(x, z)\}.$$

Stochastic Mirror Descent

Need to generalize the notion of gradient descent. Ingredients:

- A general norm $\|\cdot\|$ on $\mathbb{R}^n$
- A *distance-generating function* $\omega : X \to \mathbb{R}$ continuous and strongly convex
- Define the *prox-function* (or Bregman divergence) $V : X \times X \to \mathbb{R}$ as

$$V(x, z) = \omega(z) - \omega(x) - \nabla\omega(x)^T(z - x).$$

- For $x \in X$ define the *prox-mapping* $P_x : \mathbb{R}^n \to X$ as

$$P_x(y) = \arg\min_{x \in X}\{y^T(z - x) + V(x, z)\}.$$

Behind this hides Fenchel duality (Shalev-Shwartz, Singer 2006)...

Stochastic Mirror Descent

Example (Euclidean case): $\|\cdot\| = \|\cdot\|_2$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, we have $V(x,z) = \frac{1}{2}\|z-x\|_2^2$ and $P_x(y) = \Pi_X(x-y)$.

Stochastic Mirror Descent

Example (Euclidean case): $\|\cdot\| = \|\cdot\|_2$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, we have $V(x,z) = \frac{1}{2}\|z-x\|_2^2$ and $P_x(y) = \Pi_X(x-y)$.

In this case, SGD can be written in the form

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) \qquad \text{(SMD)}$$

.

Stochastic Mirror Descent

Example (Euclidean case): $\|\cdot\| = \|\cdot\|_2$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, we have $V(x,z) = \frac{1}{2}\|z-x\|_2^2$ and $P_x(y) = \Pi_X(x-y)$.

In this case, SGD can be written in the form

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) \qquad \text{(SMD)}$$

.

Now, forget about the Euclidean case and consider the following l_1 setup. (But keep in mind the iteration SMD).

Stochastic Mirror Descent, the l_1 setup

Few remarks: In the case of the MWU, we operate on probability distributions. X will be the probability simplex. The l_1 norm seems more natural than the l_2 norm.

Stochastic Mirror Descent, the l_1 setup

Few remarks: In the case of the MWU, we operate on probability distributions. X will be the probability simplex. The l_1 norm seems more natural than the l_2 norm.

Set $\|\cdot\| = \|\cdot\|_1$, and $\omega(x) = \sum_{i=1}^{k} x_i \log x_i$.

Stochastic Mirror Descent, the l_1 setup

Few remarks: In the case of the MWU, we operate on probability distributions. X will be the probability simplex. The l_1 norm seems more natural than the l_2 norm.

Set $\|\cdot\| = \|\cdot\|_1$, and $\omega(x) = \sum_{i=1}^{k} x_i \log x_i$. Then:

$$V(x,z) = \sum_{i=1}^{k} z_i \log \frac{z_i}{x_i}$$

and the prox-mapping is

$$[P_x(y)]_i = \frac{x_i \exp(-y_i)}{\sum_{i=1}^{k} x_i \exp(-y_i)} \qquad i = 1, \cdots, k$$

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$.

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$,

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$, and the Stochastic Mirror Descent iteration is ...

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$, and the Stochastic Mirror Descent iteration is ... the Multiplicative Weights Update :

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) = \frac{x_{t,i} \exp(-\gamma_t \xi_{t,i})}{\sum_{i=1}^{k} x_i \exp(-\gamma_t \xi_{t,i})}.$$

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$, and the Stochastic Mirror Descent iteration is ... the Multiplicative Weights Update :

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) = \frac{x_{t,i} \exp(-\gamma_t \xi_{t,i})}{\sum_{i=1}^{k} x_i \exp(-\gamma_t \xi_{t,i})}.$$

Conclusion:

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$, and the Stochastic Mirror Descent iteration is ... the Multiplicative Weights Update :

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) = \frac{x_{t,i} \exp(-\gamma_t \xi_{t,i})}{\sum_{i=1}^{k} x_i \exp(-\gamma_t \xi_{t,i})}.$$

Conclusion:

- Regret minimization can be seen as a discrete approximation of a stochastic optimization problem.

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$, and the Stochastic Mirror Descent iteration is ... the Multiplicative Weights Update :

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) = \frac{x_{t,i} \exp(-\gamma_t \xi_{t,i})}{\sum_{i=1}^{k} x_i \exp(-\gamma_t \xi_{t,i})}.$$

Conclusion:

- Regret minimization can be seen as a discrete approximation of a stochastic optimization problem.
- Multiplicative Updates algorithm is a mirror descent method on the full (non sampled) objective.

Stochastic Mirror Descent, back to experts

Recall the form of our loss function: $F(x, \xi) = x \cdot \xi$. Then $G(x, \xi) = \xi$, and the Stochastic Mirror Descent iteration is ... the Multiplicative Weights Update :

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t, \xi_t)) = \frac{x_{t,i} \exp(-\gamma_t \xi_{t,i})}{\sum_{i=1}^{k} x_i \exp(-\gamma_t \xi_{t,i})}.$$

Conclusion:

- Regret minimization can be seen as a discrete approximation of a stochastic optimization problem.
- Multiplicative Updates algorithm is a mirror descent method on the full (non sampled) objective.
- Actually, **any** online learning problem can be **optimally** solved with an online mirror descent method (Shalev-Shwartz, Singer NIPS06, Srebro, et al. NIPS11)