Is Mirror Descent a Special Case of Exponential Weights?

Dirk van der Hoeven, supervisor: dr. Tim van Erven

Leiden University

Online Convex Optimization (OCO) is a sequential prediction setting that proceeds in rounds $t = 1, \ldots, T$ in which a forecaster is to predict an unknown sequence of elements $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T \in S$, where S is a convex set. In each round the forecaster suffers a convex loss $\ell_t(\boldsymbol{w}_t) : \mathbb{R}^d \to \mathbb{R}$, which accumulates over rounds. After T rounds the performance of the forecaster is measured by the regret $\mathcal{R}_T = \sum_{t=1}^T \ell_t(\boldsymbol{w}_t) - \min_{\boldsymbol{w}} \sum_{t=1}^T \ell_t(\boldsymbol{w})$, which compares the accumulated losses to an offline minimum. Under appropriate conditions it is possible to obtain a regret that is bounded by $O(\sqrt{T})$. We focus on three algorithms that obtain such a bound: Online Gradient Descent, Mirror Descent, and Exponential Weights. In literature the Online Gradient Descent and Exponential Weights are known as special cases of Mirror Descent. However, [1] observed that Online Gradient Descent can be seen as a special case of Exponential Weights. This raises the following question: are other mirror descent algorithms also special cases of Exponential Weights?

In the OCO setting the usual approach is to approximate the loss function with a first order approximation: $\hat{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{w}_t, \nabla \ell_t(\boldsymbol{w}_t) \rangle$, where $\nabla \ell_t(\boldsymbol{w}_t)$ is the gradient of ℓ_t evaluated at $\boldsymbol{w}_t$. We then run an instance of the Mirror Descent algorithm on $\hat{\ell}$, which predicts each $\boldsymbol{w}_{t+1}$ as follows:

$$\boldsymbol{w}_{t+1} = \nabla F\Big(\nabla F^*(\boldsymbol{w}_t) - \eta \nabla \ell_t(\boldsymbol{w}_t)\Big), \tag{1}$$

where η is the learning rate, F^* is a Legendre function, and F is the convex conjugate of F^*, F. Note that ∇F and ∇F^* are inverses of each other. A special case of Mirror Descent is Online Gradient Descent, which is recovered for $F^*(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_2^2$.

Another special case of Mirror Descent is Exponential weights, which is usually run in a subsetting of OCO, Prediction With Expert Advice (PWEA). In the PWEA setting each round a set of K experts give a prediction. In each round the forecaster provides a probability distribution p_t over these experts and predicts with the mean over the expert predictions. The loss of the forecaster is then the expected loss of the experts: $\sum_{k=1}^K p_t(k) \ell_t^k$. Exponential Weights predicts the following probability distribution:

$$p_t(k) = \frac{\pi(k)\ \exp(-\eta \sum_{i=1}^{t-1} \ell_i^k)}{\sum_{k=1}^K \pi(k)\ \exp(-\eta \sum_{i=1}^{t-1} \ell_i^k)}, \tag{2}$$

where π is a prior distribution on the experts, usually taken to be uniform.

However, one can also play Exponential Weights with a non-uniform prior on a continuous set of experts, parametrized by $\boldsymbol{z}$. Let $\mathcal{E} = \{p(\boldsymbol{z}) = e^{\langle \boldsymbol{z}, \boldsymbol{\theta} \rangle - F(\boldsymbol{\theta})} K(\boldsymbol{z}) | \boldsymbol{\theta} \in \Theta\}$ be an exponential family with cumulant generating function F, sufficient statistic $\boldsymbol{z}$, and carrier $K(\boldsymbol{z})$. With expert loss function $\ell^{\boldsymbol{z}} = \langle \boldsymbol{z}, \nabla \ell_t(\boldsymbol{w}_t) \rangle$ and a prior from an exponential family we obtain the main result.

Theorem 1. *Let p_{t+1} be the Exponential Weights distribution at time $t+1$ with prior π, let the loss of expert $\boldsymbol{z}$ be $\ell_t^z = \langle \boldsymbol{z}, \nabla \ell_t(\boldsymbol{w}_t) \rangle$, let the forecasters loss be $\hat{\ell}_t = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{t+1}} [\langle \boldsymbol{z}, \nabla \ell_t(\boldsymbol{w}_t) \rangle]$, and let F be the cumulant generating function of $\mathcal{E}_\pi$. Let Mirror Descent be used with F^*, the convex conjugate of cumulant generating function F. Then the Mirror Descent algorithm is the mean of the Exponential Weights algorithm:*

$$\mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{t+1}} [\boldsymbol{z}] = \boldsymbol{w}_{t+1} = \nabla F \Big(\nabla F^*(\boldsymbol{w}_t) - \eta \nabla \ell_t(\boldsymbol{w}_t) \Big). \tag{3}$$

We provide a proof sketch. Let π be a member of an exponential family with natural parameter $\boldsymbol{\theta}_\pi$. For $\ell_i^z = \langle \boldsymbol{z}, \nabla \ell_i(\boldsymbol{w}_i) \rangle$ the update step for the EW algorithm is:

$$\begin{aligned} p_{t+1}(\boldsymbol{z}) &= \frac{\pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^t \langle \boldsymbol{z}, \nabla \ell_i(\boldsymbol{w}_i) \rangle)}{\int_{\mathbb{R}^d} \pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^t, \langle \boldsymbol{z}, \nabla \ell_i(\boldsymbol{w}_i) \rangle) d\boldsymbol{z}} \\ &= \frac{\exp(-F_\pi(\boldsymbol{\theta}_\pi) + \langle \boldsymbol{\theta}_\pi, \boldsymbol{z} \rangle - \eta \sum_{i=1}^t \langle \boldsymbol{z}, \nabla \ell_i(\boldsymbol{w}_i) \rangle) K(\boldsymbol{z})}{\exp(F_\pi(\boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t \nabla \ell_i(\boldsymbol{w}_i)) - F_\pi(\boldsymbol{\theta}_\pi))} \\ &= \exp(-F_\pi(\boldsymbol{\theta}_{p_{t+1}}) + \langle \boldsymbol{\theta}_{p_{t+1}}, \boldsymbol{z} \rangle) K(\boldsymbol{z}). \end{aligned} \tag{4}$$

Hence, p_{t+1} is a member of the exponential family with cumulant generating function F and natural parameter $\boldsymbol{\theta}_{p_{t+1}} = \boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t \nabla \ell_i(\boldsymbol{w}_i))$. Using convex conjugacy and a standard property of exponential families we find that the mean is equal to the Mirror Descent prediction:

$$\begin{aligned} \boldsymbol{\mu}_{p_{t+1}} &= \nabla F(\boldsymbol{\theta}_{p_{t+1}}) \\ &= \nabla F(\nabla F^*(\boldsymbol{w}_t) - \eta \nabla \ell_t(\boldsymbol{w}_t)) \end{aligned} \tag{5}$$

For a formal proof see Theorem 3 in [2]. Theorem 1 gives a large class of algorithms a new interpretation: the update step is updating a distribution and taking the mean of said distribution as the weights for the coming round. For instance, Online Gradient Descent is equal to Exponential weights with a Gaussian prior and predicting with the mean of the posterior. This new interpretation of Mirror Descent opens a route to find new algorithms in the OCO setting by exporting ideas from PWEA to OCO. Another possibility is sampling from the posterior, which is useful in the linear bandit optimization setting.

References

[1] Koolen, W.: Gradient descent as Exponential Weights. Blog February 21: http://blog.wouterkoolen.info/GDasEW/post.html/ (2016)

[2] Van der Hoeven, D.: Is Mirror Descent a special case of Exponential Weights? MSC Thesis. Available from: http://pub.math.leidenuniv.nl/~hoevendvander/ (2016)