Stat 928: Statistical Learning Theory **Lecture: 22**

Exponentiated Gradient Descent

Instructor: Sham Kakade

1 Exponentiated Gradient Descent

Now assume the decision space D is a d-dimensional simplex, i.e.

$$D = \{w \,|w_i \geq 0 \text{ and } ||w||_1 = 1\}$$

The Exponentiated Gradient Descent algorithm (EG) is defined as follows: at time $t = 1$, choose w_1 as the center point of this scaled simplex, namely $w_{1,i} = \frac{1}{d}$, and then use the update:

$$\forall i \in [d],\ w_{t+1,i} = \frac{w_{t,i} \exp\left(-\eta[\nabla c_t(w_t)]_i\right)}{Z_t} \tag{1}$$

where

$$Z_t = \sum_i w_{t,i}[\nabla c_t(w_t)]_i \ .$$

Here, $[\cdot]_i$ denotes the ith component of a vector. The division by Z_t serves as a form of normalization, so that $w_{t+1} \in D$, i.e. $||w_{t+1}||_1 = 1$.

We now state the guarantee of EG.

Theorem 1.1. *Assume that D is a simplex and assume that gradient is bounded as follows:*

$$||\nabla c_t(w_t)||_\infty \leq G_\infty$$

where $||u||_\infty = \max_i |u_i|$ is the L_∞ norm. If $\eta = \frac{1}{G_\infty}\sqrt{\frac{\log d}{T}}$, the regret of EG at time T bounded as:

$$R_T(EG) \leq 2G_\infty\sqrt{T \log d}$$

Now consider the decision space D to be a (scaled) d-dimensional simplex, i.e.

$$D = \{w \,|w_i \geq 0 \text{ and } ||w||_1 = D_1\}$$

EG is modified as follows: at time $t = 1$, choose w_1 as the center point of this scaled simplex, namely $w_{1,i} = \frac{D_1}{d}$, and then use the update:

$$w_{t+1,i} = \frac{w_{t,i} \exp\left(-\eta[\nabla c_t(w_t)]_i\right)}{Z_t}$$

where

$$Z_t = \frac{1}{D_1}\sum_i w_{t,i}[\nabla c_t(w_t)]_i \ .$$

Again, the division by Z serves as a form of normalization, so that $w_{t+1} \in D$, i.e. $||w_{t+1}||_1 = D_1$.

The guarantee is now:

Theorem 1.2. *Assume that D is a (scaled) simplex as defined above and assume that gradient is bounded as follows:*

$$||\nabla c_t(w_t)||_\infty \leq G_\infty$$

where $||u||_\infty = \max_i |u_i|$ is the L_∞ norm. If $\eta = \frac{1}{D_1 G_\infty}\sqrt{\frac{\log d}{T}}$, the regret of EG at time T bounded as:

$$R_T(EG) \leq 2D_1 G_\infty \sqrt{T \log d}$$

Note that the statement uses the dual norms L_1/L_∞ rather than L_2/L_2. Hence, when $D_1 G_\infty$ is $O(p)$ (where p is the number of "relevant" dimensions), this bound is only logarithmic in the total number of dimensions.

We now provide the proof using the following Lemma. The theorem follows using the learning rate specified and by verifying that the technical condition on the learning rate ($\eta \leq \frac{1}{G_\infty}$) is satisfied.

Lemma 1.3. *Let w^* be an arbitrary point in D, where D is the simplex. If $\eta \leq \frac{1}{G_\infty}$, then*

$$\sum_{t=1}^{T} \nabla_t \cdot (w_t - w^*) \leq \frac{KL(w^*||w_1)}{\eta} + \eta G_\infty^2 T \ .$$

Proof. We can interpret $w \in D$ as a probability distribution. First, it is straightforward to prove that $\exp(x) \leq 1 + x + x^2$, if $x \leq 1$. Let us examine how the KL-distance changes with respect to w^*.

$$\begin{aligned} KL(w^*||w_t) - KL(w^*||w_{t+1}) &= \sum_i w_i^* \log \frac{w_{t+1,i}}{w_{t,i}} \\ &= \sum_i w_i^*(-\eta \nabla_{t,i} - \log(Z)) \\ &= -\eta w^* \cdot \nabla_t - \log(Z) \end{aligned}$$

Now let us use that $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ to upper bound $\log(Z)$. Note that $\eta \nabla_{t,i} \leq 1$ since $\eta \leq \frac{1}{G_\infty}$.

$$\begin{aligned} \log(Z) &= \log \sum_i w_{t,i} \exp(-\eta \nabla_{t,i}) \\ &\leq \log \sum_i w_{t,i}(1 - \eta \nabla_{t,i} + \eta^2 \nabla_{t,i}^2) \\ &= \log(1 - \eta w_t \cdot \nabla_t + \eta^2 \sum_i w_{t,i} \nabla_{t,i}^2) \\ &\leq \log(1 - \eta w_t \cdot \nabla_t + \eta^2 G_\infty^2) \\ &\leq -\eta w_t \cdot \nabla_t + \eta^2 G_\infty^2 \end{aligned}$$

Combining these two we have:

$$KL(w^*||w_t) - KL(w^*||w_{t+1}) \geq -\eta w^* \cdot \nabla_t + \eta w_t \cdot \nabla_t - \eta^2 G_\infty^2$$

and so

$$\nabla_t \cdot (w_t - w^*) \leq \frac{1}{\eta}(KL(w^*||w_t) - KL(w^*||w_{t+1})) + \eta G_\infty^2$$

Summing we have:

$$\sum_{t=1}^{T} \nabla_t \cdot (w_t - w^*) \leq \frac{1}{\eta}(KL(w^*||w_1) - KL(w^*||w_{T+1})) + \eta G_\infty^2 T$$

For the uniform distribution $KL(w^*||w_1) \leq \log d$, which leads to:

$$\sum_{t=1}^{T} \nabla_t \cdot (w_t - w^*) \leq \frac{KL(w^*||w_1)}{\eta} + \eta G_\infty^2 T$$

For the case where D is a scaled simplex, we can complete the proof by rescaling by D_1. $\square$

2 Applications of Online Convex Programming

2.1 Optimization

Consider the case where we wish to optimize a convex function $c(\cdot)$ over a convex domain D. Let us run the GD algorithm, where at each time step:

$$c_t = c$$

Hence, we have the guarantee that:

$$R_T(GD) = \sum_{t=1}^{T} c(w_t) - \inf_{w \in D} \sum_{t=1}^{T} c(w) \leq D_2\, G_2 \sqrt{T}$$

where G_2 is a bound on the L_2 norm of the derivative of $c(\cdot)$.

This implies that:

$$\frac{1}{T} \sum_{t=1}^{T} c(w_t) - c(w^*) \leq \frac{D_2 G_2}{\sqrt{T}}$$

And by convexity we have:

$$c\left(\frac{1}{T} \sum_{t=1}^{T} w_t\right) \leq \frac{1}{T} \sum_{t=1}^{T} c(w_t)$$

so:

$$c\left(\frac{1}{T} \sum_{t=1}^{T} w_t\right) - c(w^*) \leq \frac{D_2 G_2}{\sqrt{T}}$$

Hence, as an optimization procedure, it is sufficient to run this algorithm for $O(\frac{1}{\epsilon^2})$ steps to get an ϵ near optimal solution.

2.2 Prediction with Expert Advice

In the ‘experts’ setting, our Decision space is $[k]$. At every round, each of the k experts provides us with a ‘suggestion’ and we choose to follow one expert. If we follow expert i at time t, we suffer loss $l_{t,i}$. As before, we do not know the loss function in advance, but once we choose our expert, we learn the full loss vector l_t.

Without a randomized strategy, it is straightforward to show that the regret must be $\omega(T)$ for some problem.

With randomization, our decision space is now a probability distribution over $[k]$.

We can view our expected loss as:

$$c_t(w) = w \cdot l_t$$

The EG algorithm, referred to as 'Hedge' for this case is: at time $t = 1$, choose w^1 as the uniform distribution, and then use the update:

$$w_{t+1} = \frac{w_t \otimes \exp(-\eta l_t)}{Z} \quad \text{where} \quad Z = w_t \cdot \exp(-\eta l_t)$$

From the guarantees of EG, we have that:

Corollary 2.1. *Assume that the losses are bounded in $[0, 1]$, i.e. $l_{t,i} \in [0, 1]$. Let w^* be an arbitrary distribution. If $\eta \leq 1$, then the expected performance of hedge is bounded as follows:*

$$\sum_{t=1}^{T} \mathbb{E}\left[l_{t,i_t}\right] - \sum_{t=1}^{T} w^* \cdot l_t \leq \frac{KL(w^* || w_1)}{\eta} + \eta T \ .$$

where i_t is random variable for the decision chosen at time t.

Hence, if we set the learning rate as $\eta = \frac{1}{\sqrt{\log dT}}$, we have that:

$$\sum_{t=1}^{T} \mathbb{E}\left[l_{t,i_t}\right] - \inf_i \sum_{t=1}^{T} \cdot l_{t,i} \leq 2\sqrt{T \log d} \ .$$

2.3 Lower Bounds

We won't formalize a lower bound. However, note that if there is the experts loss are coming from some distribution $\ell_t \sim P$, where the distribution P does not change over time, then there are distributions where it takes $\sqrt{T \log d}$ just to identify the best expert.