CSE599s, Spring 2012, Online Learning Lecture 10 - 04/26/2012

Exponentiated Gradient Descent

Lecturer: Ofer Dekel *Scribe: Albert Yu*

1 Introduction

In this lecture we review norms, dual norms, strong convexity, the Lagrange multiplier and FTRL. We look at a the Lazy Projection Gradient Descent algorithm and then develop the Exponentiated Gradient Descent algorithm.

2 Review Topics

2.1 Norms

Definition 1. *The* **norm** $\|\cdot\|$ *is a function* $\mathbb{R}^n \to \mathbb{R}$ *that gives a measure of the size of vectors in a vector space* $\mathcal{W}$*, such that* $\forall \vec{u}, \vec{v} \in \mathcal{W}$ *and scalar* a*,*

- $\|a\vec{v}\| = |a|\|\vec{v}\|$ *or positive homogeneity*
- $\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$ *or additivity (Triangle Inequality)*
- $\|\vec{v}\| \geq 0$ *and* $\|\vec{v}\| = 0 \leftrightarrow \vec{v} = \vec{0}$ *or positivity*

The norm we are primarily concerned with is the p-norm, defined as

$$\|\vec{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p}, p \geq 1 \tag{1}$$

Remark: The norm when $p = 2$ is called the Euclidean norm.

2.2 Dual Norms

Definition 2. *The* **dual norm** $\|\cdot\|^*$ *is the norm in the dual space* $\mathcal{W}^*$*, or the set of continuous linear functionals on* $\mathcal{W}$*, so that*

$$\|\vec{v}\|^* = \sup_{\|\vec{u}\| \leq 1} (\vec{u} \cdot \vec{v}) \tag{2}$$

Remark: For a given p-norm where $p \geq 1$, the dual norm is equivalent to the q-norm satsifying $1/p + 1/q = 1$. Each norm has an associated dual-norm. The Euclidean norm is a special case equal to its dual norm (or $p = q = 2$).

Lemma 3. *For a function* f *that is* G*-Lipschitz with respect to* $\|\cdot\|$*,* $\forall \vec{w}, \vec{g} \in \partial f(\vec{w})$*, then* $\|\vec{g}\|^* \leq G$

Proof. Given that f is G-Lipschitz with respect to $\|\cdot\|$

$$|f(\vec{w}') - f(\vec{w})| \leq G\|\vec{w}' - \vec{w}\| \tag{3}$$

Choose $\vec{g} \in \partial f_t(\vec{w}_t)$, where $\|\vec{g}\| \leq G$, then by (strong) convexity

$$\underbrace{|\vec{g} \cdot (\vec{w}' - \vec{w})| \leq |f(\vec{w}') - f(\vec{w})|}_{\text{Convexity}} \leq G\|\vec{w}' - \vec{w}\| \tag{4}$$

Use Hölder's Inequality were $|\vec{u} \cdot \vec{v}| \leq \|\vec{u}\|\|\vec{v}\|^*$

$$|\vec{g} \cdot (\vec{w}' - \vec{w})| \leq \|\vec{g}\|^*\|\vec{w}' - \vec{w}\| \tag{5}$$

By the definition of the dual norm presented in (2), $\|\vec{u}\| \leq 1$ and given that $\|\vec{g}\| \leq G$, then

$$|\vec{g} \cdot (\vec{w}' - \vec{w})| \leq \|\vec{g}\|^*\|\vec{w}' - \vec{w}\| \leftrightarrow \|\vec{g}\|\|\vec{w}' - \vec{w}\| \leq G\|\vec{w}' - \vec{w}\| \tag{6}$$

□

Remark: That this is roughly the same proof as stated in Lecture 4, only substituting Hölder's inequality for Cauchy-Schwarz. Note that Cauchy-Schwarz is a special case of Hölder's inequality where $p = q = 2$.

2.3 Strong Convexity (Polyak, 1966 [2])

We expand on last lecture's definition of strong convexity and offer alternative forms.

Definition 4. *A convex function, f, is* **σ-strongly convex** *with respect to some norm, $\|\cdot\|$, over a set $\mathcal{W}$ if for all $u, w \in \mathcal{W}$*

- *for every g such that $g \in \partial f(w)$ it holds that*

$$f(u) \geq f(w) + g \cdot (u - w) + \frac{\sigma}{2}\|u - w\|^2. \tag{7}$$

- *this quadratic growth in separation means that halfway between u and w*

$$f\left(\frac{u+w}{2}\right) \leq \frac{1}{2}f(u) + \frac{1}{2}f(w) - \frac{\sigma}{8}\|u - w\|^2. \tag{8}$$

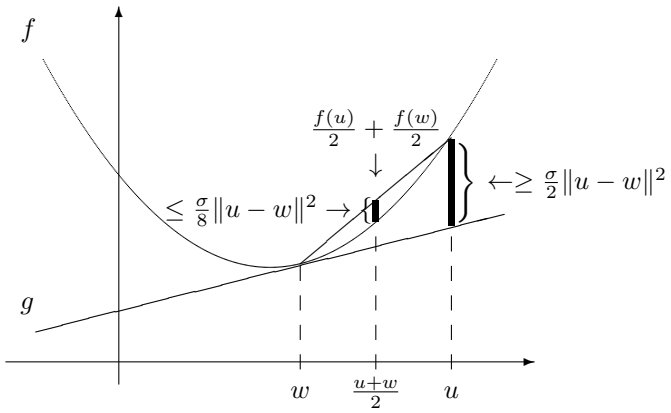

Figure 1: Depiction of strong convexity

- *(if f is twice differentiable) $\forall u$, where $\nabla^2 f$ is the Hessian of f*

$$u^T \nabla^2 f(u) u \geq \sigma \|u\|_2^2 \tag{9}$$

- *this means that the greater the step-size $(u - w)$, the greater the angle separation between gradients (in 2D the slope keeps growing), or*

$$(\nabla f(u) - \nabla f(w)) \cdot (u - w) \geq \sigma \|u - w\|^2 \tag{10}$$

Remark: Norms tend to matter more in higher dimensions, where the idea of size becomes harder to conceptualize. The above definition means that $\frac{1}{2}\|w\|_2^2$ is 1-strongly convex with respect to $\|w\|_2$, and generalized for $p \in [1, 2]$, $\frac{1}{2}\|w\|_p^2$ is $(p-1)$-strongly convex with respect to $\|w\|_p$.

2.4 Lagrange Multipliers

Lagrange Multipliers are used when we want to minimize a function subject to equality constraints via the use of a barrier function B.

1. $\min f(x)$ such that $g(x) = 0$

Define a barrier $B(x) = \begin{cases} \infty & \text{if } g(x) \neq 0 \\ 0 & \text{if } g(x) = 0 \end{cases} = \min_{\lambda \in \mathbb{R}} \lambda g(x)$, where λ is the Lagrange muliplier

2. $\min f(x) + B(x)$ such that $g(x) = 0$

Then this process becomes

3. $\min_x \max_\lambda (\underbrace{f(x) + \lambda g(x)}_{\text{Lagrangian}}) = \max_\lambda \min_x (f(x) + \lambda g(x))$

Remark: The above is under strong duality, and the Lagrange multiplier is also known as the "dual variable". The Lagrange multiplier becomes a "dialing knob" for sampling along the dominant solutions.

2.5 Follow-the-Regularized-Leader (FTRL) [3]

Previously we looked at FTRL, where at time t the player's next step with hindsight is given by

$$\vec{w}_t = \operatorname*{argmin}_{\vec{w} \in \mathcal{W}} (f_{1:(t-1)}(\vec{w}) + r(\vec{w})) \tag{11}$$

where $\mathcal{W}$ is the feasible set, $f_{1:(t-1)}(\vec{w})$ is the sum of all losses up to time $t-1$ and $r(\vec{w})$ is a strong convex regularizer (just one, not per round).

We make the following assumptions:

1. each f_t is G-Lipschitz with respect to $\|\cdot\|$ (convex)
2. $r(\vec{w})$ is σ-strongly convex with respect to $\|\cdot\|$
3. $\mathcal{W}$ is a convex set

Then we've shown in Lecture 3 that the regret for FTRL for linear f_t is bounded by

$$\text{Regret}(T) \leq \frac{TG^2}{\sigma} + r(\vec{w}^*) \tag{12}$$

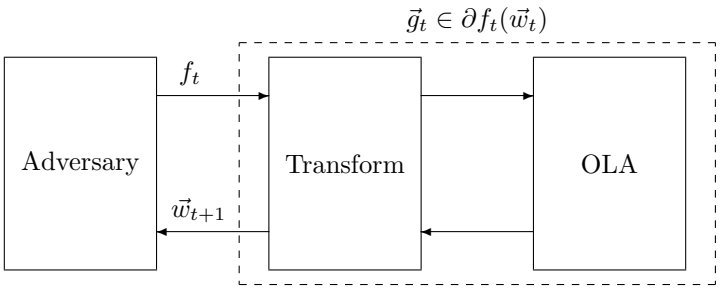

Figure 2: FTRL

Recall that linear problems are hard (worst of convex), but with a choice of $\sigma = G\sqrt{2T}/R$ we can bound the regret at $O(\sqrt{T})$.
The player can still play $\vec{w}_t = \underset{\vec{w}\in\mathcal{W}}{\operatorname{argmin}}(\vec{w} \cdot \vec{g}_{1:(t-1)} + r(\vec{w}))$ to get the bound expressed earlier in (12).

Remark: The advantage here is that the player need not keep a backlog of all previous vector $\vec{g}$, but only a running total. Each round played updates the total $\vec{g}_{1:T} \leftarrow \vec{g}_{1:(T-1)} + \vec{g}_t$. Now the question is what values of $r(\vec{w})$ should we choose?

Example: *Lazy Projection Gradient Descent Algorithm*

Choose $r(\vec{w}) = \frac{1}{2\eta}\|\vec{w}\|_2^2$, where r is $(1/\eta)$-strongly convex with respect to $\|\cdot\|_2$, and let $\eta = R/(G\sqrt{2\pi})$

- Bound: We see that our FTRL theorem still applies: $\text{Regret}(T) \leq \eta T G^2 + \frac{1}{2\eta}\|w^*\|_2^2 \leq \eta T G^2 + \frac{R^2}{2\eta}$. Choose $\eta = \frac{R}{G\sqrt{2T}}$ to get $\text{regret}(T) \leq \frac{GR\sqrt{2T}}{2} + \frac{GR\sqrt{2T}}{2} = GR\sqrt{2T}$.

- Update Rule: To find the player's next step, find $\underset{\vec{w}\in\mathcal{W}}{\operatorname{argmin}}\ \vec{w} \cdot \vec{g}_{1:(T-1)} + \dfrac{1}{2\eta}\|\vec{w}\|_2^2$. Multiply both sides by η to get (no consequence for η independent of $\vec{w}$)

$$\underset{\vec{w}\in\mathcal{W}}{\operatorname{argmin}}\ \vec{w} \cdot \vec{g}_{1:(T-1)} + \frac{1}{2\eta}\|\vec{w}\|_2^2 = \underset{\vec{w}\in\mathcal{W}}{\operatorname{argmin}}\ \vec{w} \cdot \eta\vec{g}_{1:(T-1)} + \frac{1}{2}\|\vec{w}\|_2^2 \tag{13}$$

Add a constant $\frac{1}{2}\|\eta g_{1:(t-1)}\|_2^2$ to complete the square

$$\begin{aligned} &= \underset{\vec{w}\in\mathcal{W}}{\operatorname{argmin}}\ \vec{w} \cdot \eta\vec{g}_{1:(T-1)} + \frac{1}{2}\|\vec{w}\|_2^2 \\ &= \underset{\vec{w}\in\mathcal{W}}{\operatorname{argmin}}\ \frac{1}{2}\|\eta\vec{g}_{1:(T-1)} + \vec{w}\|_2^2 \\ &= \underset{\mathcal{W}}{\operatorname{proj}}(-\eta\vec{g}_{1:(T-1)}) \end{aligned} \tag{14}$$

Remark: Abstractly, this means we see what direction our gradient points, and step in the opposite (or "downhill") direction. If our new vector is outside the feasible set, project back into it.

3 Exponentiated Gradient Descent Algorithm (EG)

3.1 What is EG

Consider a special case of FTRL where we choose

$$r(\vec{w}) = \frac{1}{\eta} \underbrace{\sum_{i=1}^{n} w_i \log w_i}_{\text{(negative) entropy}} \tag{15}$$

where w is a probability simplex such that $w \in \mathbb{R}^n : \sum_{i=1}^{n} w_i = 1, \forall i, w_i \geq 0$

Lemma 5. *For EG, r is $1/\eta$-strongly convex over $\mathcal{W}$ with respect to $\|\cdot\|_1$, so $\|w\|_1 = \sum_{i=1}^{n} |w_i|$.*

Proof. We want to show strong-convexity, which by definition (7), $\forall \vec{u}, \vec{w} \in \mathcal{W}$

$$\text{Show: } r(\vec{u}) \overset{?}{\geq} r(\vec{w}) + (\vec{u} - \vec{w}) \nabla r(\vec{w}) + \tfrac{1}{2\eta} \|\vec{u} - \vec{w}\|_1^2$$

By multiplying both sides by η and calculating $\nabla r(\vec{w})$, this is equivalent to showing

$$\Leftrightarrow \sum_{i=1}^{n} u_i \log u_i \overset{?}{\geq} \underbrace{\sum_{i=1}^{n} w_i \log w_i + \sum_{i=1}^{n} (u_i - w_i)(1 + \log w_i)}_{\sum_{i=1}^{n} u_i \log w_i} + \frac{1}{2} \|\vec{u} - \vec{w}\|_1^2 = \sum_{i=1}^{n} u_i \log w_i + \frac{1}{2} \|\vec{u} - \vec{w}\|_1^2 \tag{16}$$

Subtracting $\sum_{i=1}^{n} u_i \log w_i$ from both sides forms the inequality

$$\Leftrightarrow \sum_{i=1}^{n} u_i \log \frac{u_i}{w_i} \overset{?}{\geq} \frac{1}{2} \|\vec{u} - \vec{w}\|_1^2 \tag{17}$$

Note $\sum_{i=1}^{n} u_i \log \frac{u_i}{w_i}$ is the Kullback-Leibler divergence. By Pinsker's Inequality, $\sqrt{\frac{D(\vec{u}\|\vec{w})}{2}} \geq \sup \|\vec{u} - \vec{w}\|_1$

$$\sum_{i=1}^{n} u_i \log \frac{u_i}{w_i} = D(\vec{u}\|\vec{w}) \geq 2\|\vec{u} - \vec{w}\|_1^2 \geq \frac{1}{2} \|\vec{u} - \vec{w}\|_1^2 \tag{18}$$

with equality when $\vec{u} = \vec{w}$. This satisfies definition (7) for strong convexity. □

3.2 Update Rule

Now we look at the update rules for the given choice of $r(\vec{w})$. To find the player's next step under EG, find

$$\operatorname*{argmin}_{\vec{w} \in \mathcal{W}} \vec{w} \cdot \vec{g}_{1:(t-1)} + \frac{1}{\eta} \sum_{i=1}^{n} w_i \log w_i \tag{19}$$

Use the Lagrange multiplier and the simplex defintion for some fixed λ

$$\begin{aligned}
&= \operatorname*{argmin}_{\vec{w} \in \mathcal{W}} \vec{w} \cdot \vec{g}_{1:(t-1)} + \frac{1}{\eta} \sum_{i=1}^{n} w_i \log w_i + \lambda \left(1 - \sum_{i=1}^{n} w_i\right) \\
&= \max_{\vec{w}} \min_{\lambda} \underbrace{\left(\vec{w} \cdot \vec{g}_{1:(t-1)} + \frac{1}{\eta} \sum_{i=1}^{n} w_i \log w_i + \lambda \left(1 - \sum_{i=1}^{n} w_i\right)\right)}_{\text{Lagrangian } L} \\
&= \max_{\vec{w}} \min_{\lambda} (L)
\end{aligned} \tag{20}$$

Set the derivatives to zero and evaluate

$$\frac{\partial L}{\partial w_i} = g_{1:(t-1),i} + \frac{1}{\eta}(1 + \log w_i) - \lambda = 0 \tag{21}$$

This is equivalent to evaluating

$$\log w_i = -\eta g_{1:(t-1),i} - \underbrace{(1 - \eta\lambda)}_{\text{fixed } \lambda} \Leftrightarrow w_i = \exp\left(-\eta g_{1:(t-1),i} - (1 - \eta\lambda)\right) \tag{22}$$

Remark: Exponentiated Gradient Descent is also known as Weighted Majority, to "winnow" or to "hedge".

3.3 EG vs. GD

What are the differences between EG and GD? Consider the situation of predicting with expert advice by choosing to follow the advice of one indexed expert I_t (a random variable) on round t, out of n-experts. Let

$$g_{t,i} = \begin{cases} 1 & \text{if wrong advice by } i \\ 0 & \text{if good advice by } i \end{cases} \Rightarrow f_t = \mathsf{E}[\vec{g}_t, I_t] = \vec{g}_t \cdot \vec{w}_t \tag{23}$$

- For GD, in the all-wrong-experts scenario, $G = \sqrt{n}$ and $R(\frac{1}{2}\|w^*\|_2^2) = \frac{1}{2}$, leading to regret bound

$$\text{regret} \leq \sqrt{2Tn} \tag{24}$$

- For EG, $G = \|\vec{g}\|_\infty = \|\vec{g}\|_1^* = 1$ and $R \geq -H(w^*) = \log n$, (where H is the entropy)

$$\text{regret} \leq 2\sqrt{T \log n} \tag{25}$$

Remark: In general, EG outperforms GD, as demonstrated by their regret bounds.

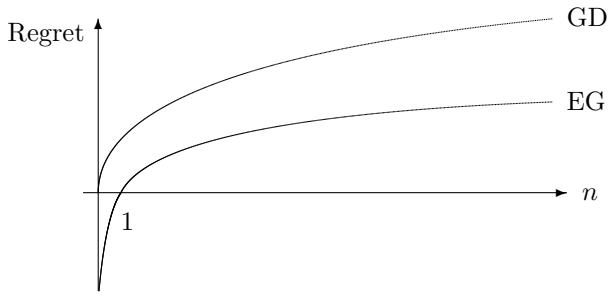

Figure 3: Relative regret bound for EG and GD with n-experts

References

[1] N. Cesa-Bianchi and G. Lugosi, "Prediction, Learning, and Games", *Cambridge University Press*, 2006.

[2] E.S. Levitin and B.T. Polyak, "Constrained Minimization Methods", *USSR Comp. Math and Math Phys.* 6, pp.1-50, 1966.

[3] H. B. McMahan, "Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L_1 Regularization", *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.