CSC 588: Machine learning theory **Spring 2021**

Lecture 20: Subgradients, Bregman divergence, Fenchel conjugate, OMD

Lecturer: Chicheng Zhang *Scribe: Erik Wessel*

1 The Power of OGD

Previously, we proved the following guarantee for online gradient descent:

Theorem 1. *For OGD initialized by $\mathbf{w}_1$ with finite step-size (a.k.a "learning rate") $\eta > 0$ the regret is bounded to be:*

$$\forall \mathbf{u} \in \Omega, R_T(\mathbf{u}) \leq \frac{\|\mathbf{u} - \mathbf{w}_1\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2$$

We also showed the following corollary when applying online-to-batch conversion:
Corollary 1 *When we have:*

- $\max_{\mathbf{u},\mathbf{v} \in \Omega} \|\mathbf{u} - \mathbf{v}\|_2 \leq B$
- $f_t(\mathbf{w}) = \ell(\mathbf{w}, \mathbf{z}_t)$ *for* $\mathbf{z}_1 ... \mathbf{z}_T \overset{iid}{\sim} \mathcal{D}$
- $f_t(\mathbf{w})$ *is ρ-Lipschitz with respect to* $\mathbf{w}_1$ *(which causes* $\|\mathbf{g}_t\|_2 \leq \rho$ *for all t)*

Then the following are guaranteed (where $\bar{\mathbf{w}} \equiv \sum_{t=1}^{T} \mathbf{w}_t$):

1. $\eta = \frac{B}{\rho}\sqrt{\frac{1}{T}} \implies \mathbb{E} L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \min_{\mathbf{w} \in \Omega} L_{\mathcal{D}}(\mathbf{w}) + \frac{B\rho}{\sqrt{T}}$
2. $\eta = \frac{1}{\rho\sqrt{T}}$, $\Omega \in \mathbb{R}^d$, $\mathbf{w}_1 = 0$, $\implies \mathbb{E} L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq L_{\mathcal{D}}(\mathbf{w}^*) + \frac{(\|\mathbf{w}^*\|_2^2 + 1)\rho}{\sqrt{T}}, \forall \mathbf{w}^* \in \mathbb{R}^d$

These guarantees are the same sort of guarantees we got for ERM earlier in the course, except the algorithm for ERM potentially had to range over entire hypothesis classes searching for the optimal model. OGD, on the other hand, *doesn't* find the ERM, instead it returns a running average of the iterant $\mathbf{w}_t$, requiring only a *single* pass over the data, and yet it achieves a similar guarantee! This is clearly a performance advantage that makes OGD very interesting, even to those uninterested in online learning applications: this simple algorithm can provide models quickly, but still achieves powerful guarantees on their generalized performance.

2 Calculating Subgradients

Now, let's introduce a procedure that is essential to convex optimization: the calculation of sub-gradients.

Helpful Fact: For $F(\mathbf{w}) = \max_{i=1}^{n}(f_i(\mathbf{w}))$, where each f_i is convex, we may define:

$$\partial F(\mathbf{w}) = \text{conv}(\cup_{i \in A(\mathbf{w})} \partial f_i(\mathbf{w}))$$

Where,

$$\text{conv}(\{x_1, ...x_n\}) = \left\{ \sum_{i=1}^{n} a_i x_i, a \in \Delta^{n-1} \right\}$$

is called the "convex combination". Δ^{n-1} is the set of vectors that are point-wise positive and sum to 1. $A(\mathbf{w})$ is defined as:

$$A(\mathbf{w}) = \{j, j \in \underset{i}{\text{argmax}}\, f_i(\mathbf{w})\}$$

For example, consider a piecewise-linear function composed of linear functions f_1, f_2, f_3, as in the following figure:

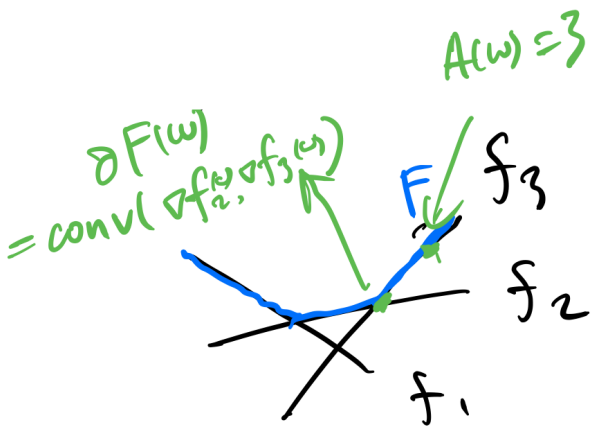

Figure 1: Subgradients of a piecewise-linear function. The black lines are the line segments, the blue line shows the piecewise function itself, and the green dots are points where we consider evaluating the subgradient sets.

At the right-most green dot, $A(\mathbf{w}) = 3$, because only f_3 is active. However, the green dot to the left is at the junction between two line segments, there both pieces are active, and we therefore get a subgradient set of: $\partial F(\mathbf{w}) = \text{conv}(\{\nabla f_2, \nabla f_3\})$.

Example: Consider the simple case $F(w) = |w|$. Here we can see that, depending on the point of evaluation, the subgradient set will be:

$$\partial F(w) = \begin{cases} \{-1\} & w < 0 \\ \text{conv}(\{-1, 1\}) & w = 0 \\ \{+1\} & w > 0 \end{cases}$$

Example: We can also consider $F(w) = \max(0, 1 - y\langle w, x\rangle)$, a.k.a the hinge loss. Then this is just a piecewise-linear function with $f_1 = 0$ and $f_2 = 1 - y\langle w, x\rangle$, and the subgradient set is:

$$\partial F(w) = \begin{cases} \{0\} & 1 - y\langle w, x\rangle < 0 \ (f_1 \text{ active}) \\ \{-\alpha yx : \alpha \in [0, 1]\} & 1 - y\langle w, x\rangle = 0 \ (\text{both active}) \\ \{-yx\} & 1 - y\langle w, x\rangle > 0 \ (f_2 \text{ active}) \end{cases}$$

3 Bregman Divergence

The notion of "distance" can be extended beyond simple norms (which were discussed previously) in a variety of ways. In this case, we will consider strongly convex functions Ψ, and define from them a generalized type of distance called the Bregman divergence.

Def: If Ψ is differentiable and strongly convex then

$$D_\Psi(\mathbf{x}, \mathbf{y}) = \Psi(\mathbf{x}) - \Psi(\mathbf{y}) - \langle \nabla\Psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$$

is the *Bregman divergence* induced by the *distance generating function* Ψ.

Graphically, the Bregman divergence can be understood as the difference between the true value of $\Psi(\mathbf{x})$ and an estimate of its value based on linear extrapolation from point $\mathbf{y}$, as shown in the figure:

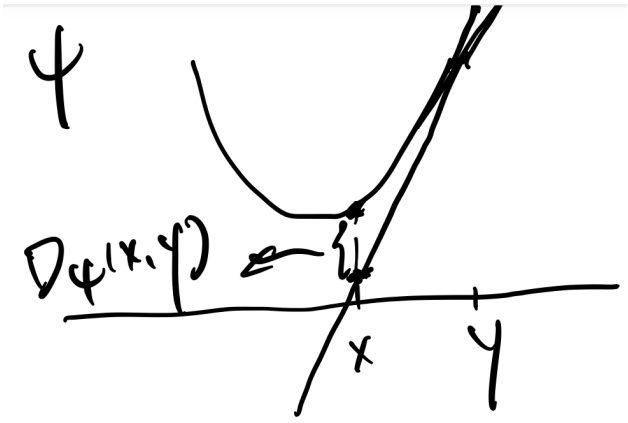

Figure 2: Using the value of $\Psi(\mathbf{y})$ and its gradient, a line extrapolates Ψ to point $\mathbf{x}$, where the extrapolation is compared with the true value $\Psi(\mathbf{x})$. The difference (always positive) is D_Ψ.

Two particularly important properties of the Bregman divergence are:

- When Ψ is λ-SC w.r.t. $\|\cdot\|$, $D_\Psi(\mathbf{x},\mathbf{y}) \geq \frac{\lambda}{2}\|\mathbf{x}-\mathbf{y}\|^2 > 0$ when $\mathbf{x} \neq \mathbf{y}$ (and is equal to 0 only when $\mathbf{x} = \mathbf{y}$).
- D_Ψ can be *asymmetric*, so $D_\Psi(\mathbf{x},\mathbf{y})$ doesn't necessarily equal $D_\Psi(\mathbf{y},\mathbf{x})$

Some of these properties are illustrated in the following important examples of Bregman divergences.
Example: $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 \implies D_\Psi(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ (using the triangle inequality).
Example: $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_A^2$ is 1-SC w.r.t. $\|\cdot\|_A = \sqrt{\mathbf{w}^\top A\mathbf{w}}$. Then $D_\Psi(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_A^2$
Example: $\Omega = \Delta^{d-1}$, $\Psi(\mathbf{w}) = \sum_{i=1}^d w_i \ln w_i$ is 1-SC w.r.t $\|\cdot\|_1$. Then $D_\Psi(\mathbf{x},\mathbf{y}) = \sum_{i=1}^d x_i \ln \frac{x_i}{y_i}$ (shown in one of the HW3 problems). This quantity is also known in information theory as "relative entropy" or the Kullback-Leibler (K-L) divergence. We can see here that it is clearly asymmetric with respect to $\mathbf{x}$ and $\mathbf{y}$.
Example: $\Psi(\mathbf{w}) = \frac{1}{2(p-1)}\|\mathbf{w}\|_p^2$ is 1-SC w.r.t. $\|\cdot\|_p$, for $p \in (1,2]$. Shalev-Shwartz (2007) compute the Bregman divergence in a long calculation (which will not be copied down here).

4 Online Mirror Descent

An important question to ask is whether we can now use these generalized norms and notions of distance to construct a more general version of gradient descent that is potentially more powerful than basic OGD. "Online mirror descent" (OMD) turns out to be such an algorithm.

We proceed as in OGD, and at each iteration we have $\mathbf{w}_t$ and a gradient taken from the subgradient set of f_t:

$$\mathbf{g}_t \in \partial f_t(\mathbf{w}_t)$$

Then, we update as follows:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}\in\Omega} \langle \mathbf{w}, \eta\mathbf{g}_t\rangle + D_\Psi(\mathbf{w},\mathbf{w_t}) \\
&= \operatorname*{argmin}_{\mathbf{w}\in\Omega} \langle \mathbf{w}, \eta\mathbf{g}_t - \nabla\Psi(\mathbf{w}_t)\rangle + \Psi(\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}\in\Omega} \langle \mathbf{w}, \nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t\rangle - \Psi(\mathbf{w})
\end{aligned}$$

If we define $\theta \equiv \nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t$, we can see that final line coincides with the definition of the *Fenchel conjugate*:

Def: If Ψ is strongly convex and Ω is a convex set, then

$$\Psi^*_\Omega(\theta) = \max_{\mathbf{w}\in\Omega} \langle \mathbf{w}, \theta\rangle - \Psi(\mathbf{w})$$

is the *Fenchel conjugate* of Ψ with respect to Ω.

Note that, since it is a pointwise maximum over linear functions, $\Psi^*_\Omega(\theta)$ is convex by construction. Graphically, the Fenchel conjugate can be thought of as being the negative of the amount by which the line defined by linear function $\langle \mathbf{w}, \theta\rangle$ would need to be moved to be tangent to the curve defined by convex function $\Psi(\mathbf{w})$, as shown:

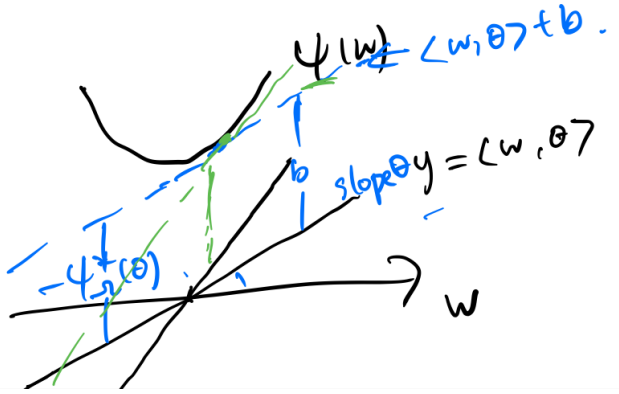

Figure 3: The black lines are $\Psi(\mathbf{w})$ and two values of $\langle \mathbf{w}, \theta\rangle$ for two different $\mathbf{w}$. The blue and green lines show the amount that each of the two black lines would have to be shifted upwards to become tangent with the $\Psi(\mathbf{w})$ curve: this is the negative of the Fenchel conjugate of Ψ w.r.t Ω.

Most importantly for understanding OMD, it can be shown that, when Ψ is convex, the gradients $\nabla\Psi^*_\Omega(\theta)$ exist $\in \mathbb{R}^d$!

$$\nabla\Psi^*_\Omega(\theta) = \operatorname*{argmax}_{\mathbf{w}\in\Omega} \langle \mathbf{w}, \theta\rangle - \Psi(\mathbf{w})$$

Therefore, we can re-write the update proceedure for OMD as:

$$\mathbf{w}_{t+1} = \nabla\Psi^*_\Omega(\nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t)$$

Graphically, this can be illustrated as:

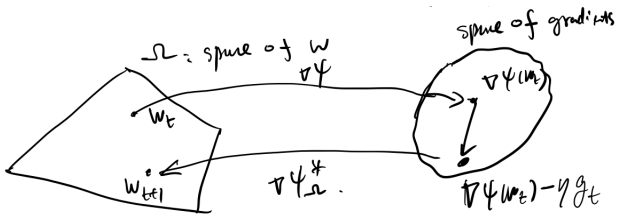

Figure 4: The process of OMD: First, we map $\mathbf{w}_t$ to the space of gradients using $\nabla\Psi$, then we execute a step of gradient descent in that space, and finally we map back to the original space with the "mirror map" $\nabla\Psi^*_\Omega$. Note that $\nabla\Psi$ and $\nabla\Psi^*_\Omega$ are not inverses of each other in general.

5 Two Important Examples of OMD

Now, we will establish two OMD-derived algorithms that are especially useful.

5.1 The p-norm algorithm

Take $\Omega = \mathbb{R}^d$, $\Psi(\mathbf{w} = \frac{1}{2}\|\mathbf{w}\|_p^2$, $p \in (1, 2]$. Then it turns out that,

$$\Psi_\Omega^*(\theta) = \frac{1}{2}\|\theta\|_q^2$$

where q is chosen such that,

$$\frac{1}{p} + \frac{1}{q} = 1$$

The derivation of this is left as an exercise, but in hinges on the fact that $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms.

It can also be shown that, in this case (Chicheng notes: only for this specific example),

$$\nabla\Psi_\Omega^* = (\nabla\Psi)^{-1}$$

where the proof of this has also been left as an exercise.

The OMD update then becomes,

$$\mathbf{w}_{t+1} = \nabla\Psi_\Omega^*(\nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t)$$

This implies,

$$\begin{aligned}
&\implies \nabla\Psi(\mathbf{w}_{t+1}) = \nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t \\
&\implies \nabla\Psi(\mathbf{w}_t) = -\eta\sum_{s=1}^{t-1}\mathbf{g}_s \\
&\implies \mathbf{w}_t = \nabla\Psi_\Omega^*\left(-\eta\sum_{s=1}^{t-1}\mathbf{g}_s\right).
\end{aligned}$$

5.2 The exponential weight algorithm

Take $\Omega = \Delta^{d-1}$, $\Psi(\mathbf{w}) = \sum_i w_i \ln w_i$, a.k.a the log-likelihood or information entropy. The initialization is $\mathbf{w}_1 = (\frac{1}{d}, \ldots\frac{1}{d})$. Then we can find the Fenchel conjugate of Ψ,

$$\begin{aligned}
\Psi_\Omega^*(\theta) &= \max_{\mathbf{w}\in\Delta^{d-1}} \langle\mathbf{w}, \theta\rangle - \sum_{i=1}^d w_i \ln w_i \\
&= \max_{\substack{w_1,\ldots w_{d-1}\geq 0 \\ \sum_{i=1}^{d-1} w_i \leq 1}} \sum_{i=1}^{d-1} w_i\theta_i + \underbrace{\left(1 - \sum_{i=1}^{d-1} w_i\right)}_{\text{Maximum } w_d}\theta_d - \sum_{i=1}^{d-1} w_i \ln w_i + \underbrace{\left(1 - \sum_{i=1}^{d-1} w_i\right)}_{\text{Maximum } w_d} \ln \underbrace{\left(1 - \sum_{i=1}^{d-1} w_i\right)}_{\text{Maximum } w_d} \\
&= \ldots \text{ (left as an exercise for HW3)} \\
&= \ln\left(\sum_{i=1}^d e^{\theta_i}\right).
\end{aligned}$$

Now we find the gradient of the Fenchel conjugate,

$$\nabla\Psi_\Omega^*(\theta) = \left(\frac{e^{\theta_i}}{\sum_{j=1}^d e^{\theta_j}}\right)_{i=1}^d$$

where the summation-style notation outside the parentheses indicate that this is a d-component vector. The gradient of the distance generating function is,

$$\nabla\Psi(\mathbf{w}) = (\ln w_i + 1)_{i=1}^{d}.$$

And so the full OMD update rule becomes,

$$\begin{aligned}
\mathbf{w}_{t+1} &= \nabla\Psi_\Omega^*(\nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t) \\
&= \nabla\Psi_\Omega^*((\ln w_{t,i} + 1 - \eta g_{t,i})_{i=1}^{d}) \\
&= \left(\frac{e^{\ln w_{t,i} + 1 - \eta g_{t,i}}}{\sum_{j=1}^{d} e^{\ln w_{t,j} + 1 - \eta g_{t,j}}}\right)_{i=1}^{d} \\
&= \left(\frac{w_{t,i} e^{-\eta g_{t,i}}}{\sum_{j=1}^{d} w_{t,j} e^{-\eta g_{t,j}}}\right)_{i=1}^{d}.
\end{aligned}$$

By induction, we can find formula that determines all $\mathbf{w}_t$ up to a constant of proportionality:

$$w_{t,i} \propto \frac{w_{1,i}}{1/d} e^{-\eta \sum_{s=1}^{t-1} g_{s,i}}.$$

And so the exact formula will be,

$$\mathbf{w}_t = \nabla\Psi_\Omega^*\left(-\eta \sum_{s=1}^{t-1} \mathbf{g}_s\right).$$

Note that in the settings of both Section 5.1 and 5.2, we both arrive at $\mathbf{w}_t = \nabla\Psi_\Omega^*\left(-\eta\sum_{s=1}^{t-1}\mathbf{g}_s\right)$. In general, this will *not* be the case. When deriving OMD updates for specific Ψ and Ω, we should always start from original OMD update $\mathbf{w}_{t+1} = \nabla\Psi_\Omega^*(\nabla\Psi(\mathbf{w}_t) - \eta\mathbf{g}_t)$.

6 Guarantees of OMD

Online mirror descent provides a generalized form of the guarantee we found for OGD:

Theorem 2. *If Ψ is 1-SC with respect to some norm $\|\cdot\|$, then OMD with distance generating function Ψ and learning rate η has regret guarantee:*

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq \frac{D_\Psi(\mathbf{u}, \mathbf{w}_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|_*^2.$$

Specifically, if $D_\Psi(\mathbf{u}, \mathbf{w}_1) \leq H^2$ and $\forall t, \|\mathbf{g}_t\|_* \leq \rho$, then:

$$\eta = \frac{H}{\rho}\sqrt{\frac{1}{T}} \implies R_T(\mathbf{u}) \leq H\rho\sqrt{T}.$$

Next lecture, this theorem will be proven.