21st Century Algorithms: Mirror Descent
Scribe Notes for 4 Oct., 2018

Roger A. Hallman
Roger.Hallman.TH@dartmouth.edu

4 October, 2018

1 Moving from Vanilla Gradient Descent to Generalized Gradient Descent

For any function f and point $x \in \mathbb{R}^n$, the dual space of $\mathbb{R}^n$, noted $(\mathbb{R}^n)^*$, is $\mathbb{R}^n$.

$$\nabla f(x) : y \mapsto \langle y, \nabla f(x) \rangle$$

indeed, ∇f at any point is a linear function.

$$\begin{aligned} \nabla f(x)(y) &= \langle y, f(x) \rangle \\ \nabla f(x)(y+z) &= \nabla f(x)(y) + \nabla f(x)(z) \\ \langle y+z, \nabla f(x) \rangle &= \langle y, \nabla f(x) \rangle + \langle z, \nabla f(x) \rangle \end{aligned}$$

Points 'live' in "Point Space" (PS) while gradients live in "Gradient Space" (GS). Consequently, gradient descent $x_{t+1} = x_t - \eta \nabla f(x_t)$ would give a "type error" because the computation involved different types. However, gradient descent is also (in our case) an isomorphic mapping. So given x_t and an identity map $y_t = I(x_t)$,

$$\begin{aligned} y_t &= I(x_t) \\ y_{t+1} &= y_t - \eta(x_t) \\ z_{t+1} &= I^{-1}(y_{t+1}) \\ x_{t+1} &= \Pi_S(z_{t+1}) = \arg\min_{v \in S} ||v - z_{t+1}|| \end{aligned}$$

> $I : PS \to GS$ is a bijection since $PS \cong GS = \mathbb{R}^n$.

The question then arises: which maps? It turns out that any differentiable function f and its gradient ∇f gives a map:

$$\begin{aligned} \Phi &: \mathbb{R}^n \to \mathbb{R} \\ I &: \underbrace{x}_{PS} \mapsto \nabla \underbrace{\Phi(x)}_{GS} \end{aligned}$$

So which $\Phi \mapsto I$ should we use? Well, if

$$\begin{aligned} \Phi(x) &= \frac{1}{2}\langle x, x \rangle = \frac{1}{2}||x||_2^2 \\ &\Rightarrow \nabla\Phi(x) = x \end{aligned}$$

Now suppose that Φ is strictly convex (i.e., $\Phi(y) > \Phi(x) + (y-x)^T \nabla\Phi(x)$). This implies that $x \mapsto \Phi(x)$ is invertable.

> Question: Can $\nabla\Phi(x) = \nabla\Phi(y)$ if $x \neq y$?

We still want $\nabla\Phi^{-1}(z)$ to exist. Suppose that Φ is $1-$strongly convex (recall the definition from our lecture on the 2nd of October, 2018: Analysis of Gradient Descent for Smooth Convex Functions. Strongly Convex and Smooth Convex Functions) and let $||\cdot||_\star$ be an arbitrary norm. With respect to $||\cdot||_\star$, i.e.,

$$\begin{aligned}
\Phi(y) &\geq \Phi(x) + (y-x)^T\nabla\Phi(x) + \frac{1}{2}||y-x||_\star^2 \\
\Phi(x) &\geq \Phi(y) + (x-y)^T\nabla\Phi(x) + \frac{1}{2}||x-y||_\star^2 \\
\implies &(y-x)^T(\nabla\Phi(y) - \nabla\Phi(x)) \geq ||y-x||_\star^2
\end{aligned}$$

Consider the following example:

$$\Phi(x) = \sum_{t=1}^{n} x_i \ln x_i$$

$\Phi : \mathbb{R}^n_{x>0} \to \mathbb{R}^n$ and $\nabla\Phi(x) = [1 + \ln x_i]$. Moreover, $\Phi^{-1}(z) = x$ and $z \in \mathbb{R}^n, z_i = 1 + \ln x_i$
$\Rightarrow x_i = e^{z_i - 1}$. We claim that Φ is $1-$strongly convex with respect to $||\cdot||_1$, i.e.,

$$\forall y, x; \Phi(y) \geq \Phi(x) + (y-x)^T\nabla\Phi(x) + \frac{1}{2}||y-x||_1^2$$

Or more precicely,

$$\sum_i y_i \ln y_i \geq \sum_i x_i \ln x_i + \sum_i [(y_i - x_i)(1 - \ln x_i)] + \frac{1}{2}\left(\sum_i |x - x_i|\right)$$

2 Bregman Divergence

Given $\Phi : \mathbb{R}^n \to \mathbb{R}$ that is strictly convex, the Bregman Divergence of Φ, $D_\Phi(y, x)$ is defined as follows:

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - (y-x)^T\nabla\Phi(x)$$

If $x \neq y$ then $D_\Phi(x, y) > 0$; note, as well, that in general $D_\Phi(x, y) \neq D_\Phi(y, x)$. If Φ is $1-$strongly convex with respect to $||\cdot||_\star$, then the following inequality holds:

$$D_\Phi(y, x) + D_\Phi(x, y) = (y-x)^T(\nabla\Phi(y) - \nabla\Phi(x)) \geq ||y-x||_\star^2$$

Returning to an example above, if $\Phi = \frac{1}{2}||x|_2^2$, then

$$\begin{aligned}
D_\Phi(y, x) &= \frac{1}{2}\left(||y||_2^2 + \frac{1}{2}||x||_2^2 - 2\langle y, x\rangle\right) \\
&= \frac{1}{2}||y-x||_2^2
\end{aligned}$$

Now let $y, x \in \Delta_n$, a simplex in $\mathbb{R}^n$, $y_i, x_i \geq 0$, and $\sum y_i = \sum x_i = 1$. If $\Phi(x) = \sum_i x_i \ln x_i$, then

$\nabla\Phi(x) = [1 + \ln x_i]$. The Bregman Divergence is as follows:

$$\begin{aligned} D_\Phi(y,x) &= \sum_i y_i \ln x_i - \sum_i x_i \ln x_i - y - x)^T [1 + \ln x_i] \\ &= \sum_i y_i \ln \frac{y_i}{x_i} \\ &= KL(y||x) \end{aligned}$$

Saying that Φ is $1-$strongly convex is equivalent to the following condition:

$$D_\Phi(y,x) + D_\Phi(x,y) \geq ||y - x||_1^2$$

Now Pinsker's Inequality gives us,

$$KL(p||q) \geq \frac{1}{2}||p - q||_1^2$$

$\Rightarrow \Phi(x) = \sum x_i \ln x_i$ is $1-$strongly convex with respect to $||\cdot||_1$.

3 Generalized Gradient Descent, aka "Mirror Descent"

Note: Need to enter graphic for mapping from Point Space to Gradient Space, and back.

The general mapping from $x_i \in \Delta \subset PS$ to $y_i \in GS$, from $y_{i+1} \in GS$ back to $z_{i+1} \in PS$ and projecting z_{i+1} onto Δ to get x_{i+1} is as follows:

1. $y_i = \nabla\Phi(x_i)$
2. $y_{i+1} = y_i - \eta\nabla\Phi(x_i)$
3. $z_{i+1} = \nabla\Phi^{-1}(y_{i+1})$
4. $x_{i+1} = \arg\min_{\rho\in\Delta} D_\Phi(\rho, z_{i+1})$

To provide illustration of this process, we return to an earlier example: $\Phi(x) = \sum x_i \ln x_i$. The mapping is as follows:

1. $[y_t]_i = 1 + \ln [x_t]_i$
2. $[y_{t+1}]_i = [y_t]_i - \eta [\nabla f(x_t)]_i = 1 + \ln [x_t]_i - \eta [\nabla f(x_t)]_i$
3. $z_{t+1} = \nabla\Phi_{-1}(y_{t+1}) = e^{\ln [x_t]_i - \eta[\nabla f(x_t)]_i} = [x_t]_i e^{-\eta[\nabla f(x_t)]_i}$
4. $[x_{t+1}]_i = \frac{[z+t+1]_i}{\sum_{i=1}^n [z_{t+1}]_i}$, so $[x_{t+1}] = \text{scaling}(z_{t+1})$

We want to $\min_{x\in S} f(x)$.
Consider mirror descent for $P(x) = \sum_i x_i \ln x_i$ gives us

$$(x_{t+1})_i = \text{scaling}\left(x_t(i) s^{-\eta[\nabla f(x_t)]_i}\right)$$

whereas "vanilla gradient descent" gives us

$$(x_{t+1})_i = \text{"projection } \Delta\text{"}\ (x_t(i) - \eta [\nabla f(x_t)]_i)$$

Our analysis follows the same approach as earlier:

$$
\begin{aligned}
f(x_t) - f(x_\star) = \text{err}(t) &\leq (x_t - x_\star)^T \nabla f(x_t) && \text{(because of the convexity of } f\text{)} \\
&= \frac{1}{\eta}(x_t - x_\star)^T (y_t - y_\star) && \text{(the mirror descent algorithm)} \\
&= \frac{1}{\eta}(x_t - x_\star)^T (\nabla\Phi(x_t) - \nabla\Phi(z_{t+1}) \\
&= \frac{1}{\eta}(x_t - x_\star)^T (\nabla\Phi(z_{t+1}) - \nabla\Phi(x_t)) \\
&= \frac{1}{\eta}\left(D_\Phi(x_\star, x_t) + D_\Phi(x_t, z_{t+1}) - D_\Phi(x_\star, z_{t+1})\right) && \text{Bregman Cosine} \\
&\leq \frac{1}{\eta}\left(D_\Phi(x_\star, x_t) + D_\Phi(x_t, z_{t+1}) - D_\Phi(x_\star, x_{t+1})\right) \\
\text{Now take } D_t &= D_\Phi(x_\star, x_t) \\
&= \frac{1}{\eta}\underbrace{(D_t - D_{t+1})}_{\text{telescopes}} + \frac{1}{\eta}(x_t, z_{t+1})
\end{aligned}
$$

Properties of the Bregman Divergence

1. The Bregman Cosine Rule: $\langle u - v, \nabla\Phi(w) - \nabla\Phi(v)\rangle = D_\Phi(u, v) + D_\Phi(v, u) - D_\Phi(u, w)$
2. The Bregman Projection: $\forall u \in S, v \notin S, w \in \Pi_S^\Phi(v)$, $D_\Phi(u, v) \geq D_\Phi(u, w) + D_\Phi(w, v)$
3. $D_\Phi(u, v) + D_\Phi(v, u) = \langle u - v, \nabla\Phi(u) - \nabla\Phi(v)\rangle$
4. $D_\Phi(u, v) \geq \frac{1}{2}||u - v||^2$ (i.e., $1-$strong convexity)
   $\Rightarrow D_\Phi(v, u) \leq \langle u - v, \nabla\Phi(u) - \nabla\Phi(v)\rangle - \frac{1}{2}||u - v||^2$

Now,

$$
\begin{aligned}
D_\Phi(x_t, z_{t+1}) &\leq \langle x_t - z_{t+1}, \nabla\Phi(x_t) - \nabla\Phi(z_{t+1})\rangle - \frac{1}{2}||x_t - z_{t+1}||^2 \\
&= \eta\langle x_t - z_{t+1}, \nabla f(x_t)\rangle - \frac{1}{2}||x_t - z_{t+1}||^2 \\
&\leq \eta||x_t - z_{t+1}|| \cdot ||\nabla f(x_t)||_\infty - \frac{1}{2}||x_t - z_{t+1}||^2 \\
&\leq \frac{\eta^2}{2}||\nabla f(x_t||_\infty^2 \\
&\leq \frac{\eta^2\rho^2}{2}
\end{aligned}
$$

Finally, we come to the following theorem:

To minimize $f(x)$ over Δ_n, if $||\nabla f(x)||_\infty \leq \rho$ and $D_\Phi(x_1, x_\star) \leq D$, takes $\frac{D\rho^2}{\varepsilon^2}$ iterations.