Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Chao Qin
Email: cq2199@columbia.edu

SCRIBE
21

ANALYSIS OF THE MIRROR DESCENT ALGORITHM

In this lecture, we analyze the mirror descent algorithm.

Algorithm 1: Mirror Descent

1. Start with arbitrary $\mathbf{x}_0 \in K$
2. For $t = 0, 1, \ldots, T-1$,
   (a) Set $\mathbf{y}_{t+1} = [\nabla\phi]^{-1}\left(\nabla\phi(\mathbf{x}_t) - \eta\nabla f(\mathbf{x}_t)\right)$
   (b) Set $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in K} B_\phi(\mathbf{x}, \mathbf{y}_{t+1})$
3. Output $\bar{\mathbf{x}} \triangleq \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_t$

Assume we have the following

- Norm $\|\cdot\|$ on $\mathbb{R}^d$
- $\nabla\phi : U \to \mathbb{R}^d$ is surjective
- 1-strong convexity: $\phi(\mathbf{y}) \geq \phi(\mathbf{x}) + \nabla\phi(\mathbf{x})\cdot(\mathbf{y}-\mathbf{x}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Bregman divergence: $B_\phi(\mathbf{y}, \mathbf{x}) = \phi(\mathbf{y}) - \left(\phi(\mathbf{x}) + \nabla\phi(\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})\right)$
- $\forall \mathbf{x} \in K, \left\|\nabla f(\mathbf{x})\right\|_* \leq L$
- $\forall \mathbf{x}, \mathbf{x}' \in K, \left\|\mathbf{x}-\mathbf{x}'\right\| \leq D$

With these, we begin to analyze the mirror descent algorithm.

$$\begin{aligned}
B_\phi(\mathbf{x}^*, \mathbf{y}_{t+1}) &= \phi(\mathbf{x}^*) - \phi(\mathbf{y}_{t+1}) - \nabla\phi(\mathbf{y}_{t+1})\cdot(\mathbf{x}^* - \mathbf{y}_{t+1}) \\
&= \phi(\mathbf{x}^*) - \phi(\mathbf{x}_t) + \phi(\mathbf{x}_t) - \phi(\mathbf{y}_{t+1}) - \nabla\phi(\mathbf{y}_{t+1})\cdot(\mathbf{x}^* - \mathbf{x}_t + \mathbf{x}_t - \mathbf{y}_{t+1}) \\
&= \phi(\mathbf{x}^*) - \phi(\mathbf{x}_t) - \nabla\phi(\mathbf{y}_{t+1})\cdot\left(\mathbf{x}^* - \mathbf{x}_t\right) + B_\phi(\mathbf{x}_t - \mathbf{y}_{t+1}) \\
&= \phi(\mathbf{x}^*) - \phi(\mathbf{x}_t) - \left(\nabla\phi(\mathbf{x}_t) - \eta\nabla f(\mathbf{x}_t)\right)\cdot(\mathbf{x}^* - \mathbf{x}_t) + B_\phi(\mathbf{x}_t, \mathbf{y}_{t+1}) \\
&= B_\phi(\mathbf{x}^*, \mathbf{x}_t) + B_\phi(\mathbf{x}_t, \mathbf{y}_{t+1}) + \eta\nabla f(\mathbf{x}_t)\cdot(\mathbf{x}^* - \mathbf{x}_t) \qquad (1)
\end{aligned}$$

Lemma 21.0.1 $B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) \le B_\phi(\mathbf{x}^*, \mathbf{y}_{t+1})$.

Proof.

$$\begin{aligned} B_\phi(\mathbf{x}^*, \mathbf{y}_{t+1}) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) &= \phi(\mathbf{x}_{t+1}) - \phi(\mathbf{y}_{t+1}) - \nabla\phi(\mathbf{y}_{t+1}) \cdot (\mathbf{x}^* - \mathbf{y}_{t+1}) + \nabla\phi(\mathbf{x}_{t+1}) \cdot (\mathbf{x}^* - \mathbf{x}_{t+1}) \\ &= B_\phi(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) + (\nabla\phi(\mathbf{x}_{t+1}) - \nabla\phi(\mathbf{y}_{t+1}))(\mathbf{x}^* - \mathbf{x}_{t+1}) \\ &\ge (\nabla\phi(\mathbf{x}_{t+1}) - \nabla\phi(\mathbf{y}_{t+1})) \cdot (\mathbf{x}^* - \mathbf{x}_{t+1}) \end{aligned}$$

Notice that $B_\phi(\mathbf{x}, \mathbf{y}_{t+1})$ is convex in $\mathbf{x}$. Since $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in K} B_\phi(\mathbf{x}, \mathbf{y}_{t+1})$, by the first order optimality condition, we have

$$\nabla_{\mathbf{x}=\mathbf{x}_{t+1}} B_\phi \cdot (\mathbf{x}, \mathbf{y}_{t+1})(\mathbf{x}^* - \mathbf{x}_{t+1}) \ge 0 \iff (\nabla\phi(\mathbf{x}_{t+1}) - \nabla\phi(\mathbf{y}_{t+1})) \cdot (\mathbf{x}^* - \mathbf{x}_{t+1}) \ge 0$$

This completes the proof. □

Lemma 21.0.2 $B_\phi(\mathbf{x}_t, \mathbf{y}_{t+1}) \le \frac{\eta^2}{2} \| \nabla f(\mathbf{x}_t) \|_*^2$.

Proof.

$$\begin{aligned} B_\phi(\mathbf{x}_t, \mathbf{y}_{t+1}) + B_\phi(\mathbf{y}_{t+1}, \mathbf{x}_t) &= (\nabla\phi(\mathbf{x}_t) - \nabla\phi(\mathbf{y}_{t+1})) \cdot (\mathbf{x}_t - \mathbf{y}_{t+1}) \\ &= \eta\nabla f(\mathbf{x}_t) \cdot (\mathbf{x}_t - \mathbf{y}_{t+1}) \\ &\le \| \eta\nabla f(\mathbf{x}_t) \|_* \| \mathbf{x}_t - \mathbf{y}_{t+1} \| \\ &\le \frac{\| \eta\nabla f(\mathbf{x}_t) \|_*^2 + \| \mathbf{x}_t - \mathbf{y}_{t+1} \|^2}{2} \end{aligned}$$

where the first inequality uses the Cauchy–Schwarz inequality, and the last inequality follows from the AM–GM inequality. Notice that by 1-strongly convexity of ϕ,

$$B_\phi(\mathbf{y}_{t+1}, \mathbf{x}_t) \ge \frac{\| \mathbf{y}_{t+1} - \mathbf{x}_t \|^2}{2}.$$

This completes the proof. □

With these lemmas, Equation (1) leads to

$$B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) \le B_\phi(\mathbf{x}^*, \mathbf{x}_t) + \frac{\eta^2}{2} \| \nabla f(\mathbf{x}_t) \|_*^2 + \eta\nabla f(\mathbf{x}_t) \cdot (\mathbf{x}^* - \mathbf{x}_t).$$

Then we have

$$\eta \sum_{t=0}^{T-1} \nabla f(\mathbf{x}_t) \cdot (\mathbf{x}_t - \mathbf{x}^*) \le B_\phi(\mathbf{x}^*, \mathbf{x}_0) + \frac{\eta^2}{2} \sum_{t=0}^{T-1} \| \nabla f(\mathbf{x}_t) \|_*^2$$

Notice that

$$\text{LHS} \ge \sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^*)\big) \ge Tf(\bar{\mathbf{x}}) - Tf(\mathbf{x}^*)$$

where the first inequality uses the convexity of f and the other one follows from the Jensen's inequality. Hence,

$$f(\bar{\mathbf{x}}) - f(\mathbf{x}^*) \le \frac{B_\phi(\mathbf{x}^*, \mathbf{x}_0)}{\eta T} + \frac{\eta}{2T} \sum_{t=0}^{T-1} \| \nabla f(\mathbf{x}_t) \|_*^2 \le \frac{B_\phi(\mathbf{x}^*, \mathbf{x}_0)}{\eta T} + \frac{\eta L^2}{2} = \sqrt{\frac{2B_\phi(\mathbf{x}^*, \mathbf{x}_0)L^2}{T}}$$

by letting $\eta = \sqrt{\frac{2B_\phi(\mathbf{x}^*, \mathbf{x}_0)}{L^2 T}}$. Indeed $B_\phi(\mathbf{x}^*, \mathbf{x}_0)$ is unknown, we need to derive an upper bound of it with a particular choice of $\mathbf{x}_0$. Let $\mathbf{x}_0 \triangleq \arg\min_{\mathbf{x}\in K} \phi(\mathbf{x})$. By the first order optimality condition,

$$\nabla\phi(\mathbf{x}_0)\cdot(\mathbf{x}^*-\mathbf{x}_0)\geq 0,$$

and then

$$B_\phi(\mathbf{x}^*,\mathbf{x}_0)=\phi(\mathbf{x}^*)-\phi(\mathbf{x}_0)-\nabla\phi(\mathbf{x}_0)\cdot(\mathbf{x}^*-\mathbf{x}_0)\leq\phi(\mathbf{x}^*)-\phi(\mathbf{x}_0)\leq\max_{\mathbf{x}\in K}\phi(\mathbf{x})-\min_{\mathbf{x}\in K}\phi(\mathbf{x}).$$

0.1 $\ell 1$ Case

When minimizing a function f over the simplex Δ_d, we can choose the negative entropy function as the mirror map, i.e.,

$$\phi(\mathbf{x})=\sum_{i=1}^{d}x_i\log\left(x_i\right).$$

It is easy to check this mirror map satisfies the conditions required by the above result. Since $\max_{\mathbf{x}\in\Delta_d}\phi(\mathbf{x})-\min_{\mathbf{x}\in\Delta_d}\phi(\mathbf{x})=\log\left(d\right)$, the mirror descent algorithm achieves a rate of convergence of order $\sqrt{\frac{\log(d)}{T}}$, while the (projected) gradient descent algorithm only has a rate of order $\sqrt{\frac{d}{T}}$ in this case.

0.2 AdaGrad

The idea of mirror descent is also used to design the so-called AdaGrad algorithm. The suggested reading CEH [Chapter 6] covers AdaGrad in detail (this is outside the syllabus for the final exam however).