EE C227C Convex Optimization and Approximation Lecture 13 - 3/1/2022

Lecture 13: Accelerated Gradient Descent

Lecturer: Jiantao Jiao *Scribe: Thomas Fork and Junhao (Bear) Xiong*

1 Recap

1.1 Problem Setup

Recall the problem setup of gradient descent from the previous lecture. We consider unconstrained minimization of a function $f : \mathbb{R}^n \to \mathbb{R}$ which is smooth, convex, and has L-Lipschitz gradient with respect to an arbitrary norm $\|\cdot\|$, i.e.

$$\forall x, y \in \mathbb{R}^n, \|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \tag{1}$$

where $|\cdot\|_*$ is the dual norm associated with $\|\cdot\|$. Furthermore we are given a σ-strongly convex regularizer $R : \mathbb{R}^n \to \mathbb{R}$ whose gradient ∇R is a bijection. Recall the definition of σ-strongly convex:

$$D_R(x, y) = R(x) - R(y) - \langle \nabla R(y), x - y \rangle \geq \frac{\sigma}{2} \|x - y\|^2 \tag{2}$$

It was previously stated that under these assumptions there exists an algorithm, which when provided

1. First order oracle access
2. Access to ∇R and $(\nabla R)^{-1}$
3. $D_R(x^*, x_0) \leq D^2$
4. $\epsilon \geq 0$

produces a point $x \in \mathbb{R}^n$ such that

$$f(x) \leq f(x^*) + \epsilon$$

with $T \sim \mathbb{O}\left(\sqrt{\frac{LD^2}{\sigma\epsilon}}\right)$ queries to the oracle and $\mathbb{O}(nT)$ arithmetic operations.

1.2 Estimate Sequence

Definition 1. *A sequence $(\phi_t, \lambda_t, x_t)_{t\geq 0}$ with $\phi : \mathbb{R}^n \to \mathbb{R}$, $\lambda_t \in [0, 1]$ and $x_t \in \mathbb{R}^n$ is said to be an estimate sequence of $f : \mathbb{R}^n \to \mathbb{R}$ if the following properties hold:*

1. ***Lower Bound Property***

$$\forall t \geq 0, x \in \mathbb{R}^n, \ \phi_t(x) \leq (1 - \lambda_t) f(x) + \lambda_t \phi_0(x)$$

2. ***Upper Bound Property***

$$\forall x \in \mathbb{R}^n, \ \phi_t(x) \geq f(x_t) \Rightarrow \min_{x \in \mathbb{R}^n} \phi_t(x) \geq f(x_t)$$

In the previous lecture we proved the following regarding estimate sequences:

Lemma 2. *Under the given problem setup if we have an estimate sequence (ϕ_t, λ_t, x_t) with $\phi_0(x) = f(x_0) + \frac{L}{2\sigma} D_R(x, x_0)$ then for some constant $C > 0$*

$$f(x_t) \leq f(x^*) + \underbrace{\frac{CLD^2}{\sigma t^2}}_{=\epsilon}$$

However, it remains to demonstrate how to obtain a valid estimate sequence and obtain this bound.

1.3 Estimate Sequence Update

The previous lecture proposed the following update for ϕ_t:

$$\phi_t(x) = (1-\gamma_t)\phi_{t-1}(x) + \gamma_t L_{t-1}(x) \tag{3a}$$

$$L_{t-1}(x) = f(y_{t-1}) + \langle \nabla f(y_{t-1}), x - y_{t-1} \rangle \tag{3b}$$

without specifying the choice of γ_t and y_t. In this lecture we obtain these by enforcing the lower and upper bound properties on updates for ϕ_t, λ_t and x_t to obtain an estimate sequence.

2 Completing The Estimate Sequence Update

We enforce the upper and lower bounds on the proposed estimate sequence update rule, and demonstrate that $\phi_t(x)$ has a closed-form representation. For brevity, henceforth $L = \sigma = 1$.

2.1 Enforcing the Lower Bound

We prove this by induction. Consider again the proposed estimate sequence update:

$$\phi_t(x) = (1-\gamma_t) \underbrace{\phi_{t-1}(x)}_{\leq (1-\lambda_{t-1})f(x) + \lambda_{t-1}\phi_0(x)} + \gamma_t \underbrace{L_{t-1}(x)}_{\leq f(x)} \tag{4}$$

substituting the inequalities (from convexity and enforcing the lower bound property on ϕ_{t-1}) and rearranging terms we have

$$\phi_t(x) \leq \underbrace{((1-\gamma_t)(1-\lambda_{t-1}) + \gamma_t)}_{(1-\lambda_t)} f(x) + \underbrace{(1-\gamma_t)\lambda_{t-1}}_{\lambda_t} \phi_0(t) \tag{5}$$

which is identical in form to applying the lower bound property to $\phi_t(x)$. We force the estimate sequence update to satisfy the lower bound property by setting

$$\lambda_t = (1-\gamma_t)\lambda_{t-1} \tag{6}$$

In general, we must choose $\lambda_0 = 1$. With this, we obtain the second equation for λ_t

$$\lambda_t = \prod_{1 \leq i \leq t} (1-\gamma_i) \tag{7}$$

Restrictions on the choice of γ_t and a specific example will be provided later.

2.2 Representation of $\phi_t(x)$

Next we prove that we can represent $\phi_t(x)$ as $\phi_t(x) = \phi_t^* + \lambda_t D_R(x, z_t)$ and perform the proposed update while keeping the same representation. Here ϕ_t^* is a constant, and by definition the minimum of ϕ_t, which occurs at z_t, since by definition $\phi_t(x) = \phi_t^* + \lambda_t D_R(x, z_t)$, and $D_R(x, z_t) = 0$ when $x = z_t$. We use the following lemma regarding a property for Bregman divergence:

Lemma 3. ***(Bregman Shifting)*** *Let $z \in \mathbb{R}^n$ and R be a suitable convex regularizer such that ∇R is a bijection.*

Then

$$\forall l \in \mathbb{R}^n, \exists z' \in \mathbb{R}^n : \begin{array}{c} D_R(x,z) + \langle x - z, l \rangle = D_R(x,z') - D_R(z,z') \\ l = \nabla R(z) - \nabla R(z') \end{array}$$

Now we use this to prove closedness of the representation of $\phi_t(x)$ by induction

Suppose $z_0 = x_0$, $\lambda_t = (1-\gamma_t)\lambda_{t-1}$ and $\phi_{t-1}(x) = \phi^*_{t-1} + \lambda_{t-1}D_R(x, z_{t-1})$ then the proposed estimate sequence update is (using the inductive hypothesis on $\phi_{t-1}(x)$)

$$\phi_t(x) = (1-\gamma_t)\left(\phi^*_{t-1} + \lambda_{t-1}D_R(x, z_{t-1})\right) + \gamma_t\left[f(y_{t-1}) + \langle x - y_{t-1}, \nabla f(y_{t-1})\rangle\right] \tag{8}$$

$$= (1-\gamma_t)\phi^*_{t-1} + \underbrace{\lambda_t D_R(x, z_{t-1}) + \gamma_t\langle x, \nabla f(y_{t-1})\rangle}_{\text{only these terms are not constant}} + \gamma_t\left[f(y_{t-1}) - \langle y_{t-1}, \nabla f(y_{t-1})\rangle\right] \tag{9}$$

focusing on the terms that are not constant and using the Bregman Shifting Lemma with $l = \frac{\gamma_t}{\lambda_t}\nabla f(y_{t-1})$, $z = z_{t-1}$ and $z' = z_t$:

$$= \lambda_t\left(D_R(x, z_{t-1}) + \frac{\gamma_t}{\lambda_t}\langle x, \nabla f(y_{t-1})\rangle\right) \tag{10}$$

$$= \lambda_t\left(D_R(x, z_t) - D_R(z_{t-1}, z_t) + \left\langle z_{t-1}, \frac{\gamma_t}{\lambda_t}\nabla f(y_{t-1})\right\rangle\right) \tag{11}$$

$$= \lambda_t\left(D_R(x, z_t) + \text{constant}\right) \tag{12}$$

therefore the representation $\phi_t(x) = \phi^*_t + \lambda_t D_R(x, z_t)$ is a closed form for $\phi_t(x)$ under the proposed update rule. Furthermore, z_t is automatically the minimizer of ϕ_t. All of the terms that are constant are captured by the updated ϕ^*_t.

The above proof also gives us an update rule for z_t from the Bregman Shifting Lemma:

$$\nabla R(z_t) = \nabla R(z_{t+1}) - \frac{\gamma_t}{\lambda_t}\nabla f(y_{t-1}) \tag{13}$$

2.3 Enforcing the Upper Bound

It remains to enforce the Upper Bound Property on our estimate sequence. Before proceeding, it is worth noting our current progress:

- We have a relationship for $\lambda_{t-1} \to \lambda_t$ using γ_t, but no restrictions on γ_t.
- We have a relationship for $z_{t-1} \to z_t$.
- We don't yet have update rules for y_t and x_t.

Recall the update rule

$$\phi_t(x) = (1-\gamma_t)\left(\phi^*_{t-1} + \lambda_{t-1}D_R(x, z_{t-1})\right) + \gamma_t\left[f(y_{t-1}) + \langle x - y_{t-1}, \nabla f(y_{t-1})\rangle\right] \tag{14}$$

We can remove ϕ^*_{t-1} using the inductive hypothesis of the upper bound property, and then invoke strong convexity:

$$\phi_t(x) \geq \underbrace{(1-\gamma_t)f(x_{t-1})}_{\geq (1-\gamma_t)(f(y_{t-1}) + \langle x_{t-1} - y_{t-1}, \nabla f(y_{t-1})\rangle)} + \lambda_t D_R(x, z_{t-1}) + \gamma_t\left(f(y_{t-1}) + \langle x - y_{t-1}, \nabla f(y_{t-1})\rangle\right) \tag{15}$$

$$= f(y_{t-1}) + \lambda_t D_R(x, z_{t-1}) + \gamma_t\langle x - z_{t-1}, \nabla f(y_{t-1})\rangle + \left\langle \nabla f(y_{t-1}), \underbrace{(1-\gamma_t)x_{t-1} + \gamma_t z_{t-1} - y_{t-1}}_{\text{We set this to 0}}\right\rangle \tag{16}$$

where we add and subtract the term $\gamma_t\langle x - z_{t-1}, \nabla f(y_{t-1})\rangle$. Observe that setting

$$y_{t-1} = (1-\gamma_t)x_{t-1} + \gamma_t z_{t-1} \tag{17}$$

brings the underbraced term to 0. This choice is not immediately obvious, however it leaves us with the form

$$\phi_t(x) \geq f(y_{t-1}) + \lambda_t D_R(x, z_{t-1}) + \gamma_t \langle x - z_{t-1}, \nabla f(y_{t-1}) \rangle \tag{18}$$

or using 1-strong convexity of R

$$\phi_t(x) \geq f(y_{t-1}) + \gamma_t \langle x - z_{t-1}, \nabla f(y_{t-1}) \rangle + \frac{\lambda_t}{2} \|x - z_{t-1}\|^2 \tag{19}$$

which is very close to the L smoothness property enforced on f.

Define $\tilde{x} - y_{t-1} = \gamma_t(x - z_{t-1})$. Then:

$$\phi_t(x) \geq f(y_{t-1}) + \langle \tilde{x} - y_{t-1}, \nabla f(y_{t-1}) \rangle + \frac{\lambda_t}{2\gamma_t^2} \|\tilde{x} - y_{t-1}\|^2 \tag{20}$$

if we enforce the restriction (intuitively, this also tells us λ_t can't be too small)

$$\frac{\lambda_t}{\gamma_t^2} \geq 1 \tag{21}$$

then

$$\phi_t(x) \geq f(y_{t-1}) + \langle \tilde{x} - y_{t-1}, \nabla f(y_{t-1}) \rangle + \frac{1}{2} \|\tilde{x} - y_{t-1}\|^2 \geq f(\tilde{x}) \tag{22}$$

i.e. $\phi_t(x)$ is lower bounded by an upper bound on f. The best x_t we can choose for this upper bound is

$$x_t \triangleq \arg\min_{\tilde{x}} \langle \tilde{x}, \nabla f(y_{t-1}) \rangle + \frac{1}{2} \|\tilde{x} - y_{t-1}\|^2 \tag{23}$$

This provides us with update rules for y_t and x_t, and a limit on our choice of γ_t, all of which now satisfy the lower and upper bound propertyes for a valid estimate sequence.

3 The Accelerated Gradient Algorithm

Using $z_0 = x_0$, $\lambda_0 = 1$, $\frac{\lambda_t}{\gamma_t^2} \geq 1$ and $\lambda_t = \prod_{1 \leq i \leq t}(1 - \gamma_t)$ we have

$$y_{t-1} = (1 - \gamma_t)x_{t-1} + \gamma_t z_{t-1} \tag{24a}$$

$$\nabla R(z_t) = \nabla R(z_{t-1}) - \frac{\gamma_t}{\lambda_t} \nabla f(y_{t-1}) \tag{24b}$$

$$x_t = \arg\min_{\tilde{x}} \langle \tilde{x}, \nabla f(y_{t-1}) \rangle + \frac{1}{2} \|\tilde{x} - y_{t-1}\|^2 \tag{24c}$$

3.1 Observations

Notice that y_t is a linear combination of x_t and z_t, each of which are updated separately: x_t follows the update of gradient descent wheras z_t follows the update of mirror descent. We have not yet shown that we can pick γ_t such that $\lambda_t \approx \frac{1}{t^2}$ so we have not yet shown the desired convergence rate of

$$f(x_t) \leq f^* + O\left(\frac{LD^2}{\sigma t^2}\right) \tag{25}$$

3.2 Choosing γ_t

We previously made the assumption $\lambda_t \geq \gamma_t^2$, which limits our ability to make λ_t decay. Consider the choice of γ_t from [1, Ch. 8.5.1]:

$$\gamma_t = \begin{cases} 0 & t \in \{0,1,2,3\} \\ \frac{2}{t} & \text{otherwise} \end{cases} \tag{26}$$

It follows that

$$\lambda_t = \begin{cases} 1 & t \in \{0,1,2,3\} \\ \frac{6}{t(t-1)} & \text{otherwise} \end{cases} \tag{27}$$

Since $\frac{6}{t(t-1)} \geq \frac{4}{t^2} = \gamma^2$ this is a valid sequence of γ_t and λ_t

Having found a valid sequence of λ_t this completes the derivation and performance bound of the accelerated gradient algorithm.

References

[1] N. K. Vishnoi, *Algorithms for Convex Optimization.* Cambridge University Press, 2020. [Online]. Available: https://convex-optimization.github.io/ACO-v1.pdf