Lecture 12: Optimization Algorithms

Date: 03/15/21

Lecturer: Andre Wibisono *Scribing by: Jon Patsenker*

12.1. Optimization through continuous time

We would like to be able to minimize some objective function over a search space. For our purposes we will be solving,

$$\min_{x \in \mathbb{R}^n} f(x)$$

As we will be solving this problem by means of optimization, we will define a gradient flow corresponding to an optimizing process: $\dot{X}_t = -\nabla f(X_t)$.

We will assume a setting where f is differentiable, and $\min f = \min\limits_{x \in \mathbb{R}^n} f(x)$

We will also define the following properties:

- f is α - **strongly convex** if $\forall y$,

  $$f(y) \geq f(x) + \langle \nabla f,\, y - x \rangle + \frac{\alpha}{2} \|x - y\|^2$$

  which can be re-expressed as

  $$(f(x) - f(y))^T (x - y) \geq \alpha \|x - y\|^2$$

  because this is defined $\forall y$ this is equivalent to the statement

  $$\nabla^2 f(x) \succeq \alpha I$$

- f is α - **gradient dominated** if $\|\nabla f(x)\|^2 \geq 2\alpha(f(x) - \min f)$. Note that f being strongly convex implies it is gradient dominated.

12.1.1. Convergence of in continuous setting

In the continuous setting we can evaluate the stationary distributions given functions that fit one or both of these criteria.

Theorem 12.1.1. *If f is differentiable, and α-strongly convex, the gradient flow $\dot{X}_t = -\nabla f(X_t)$ exhibits exponential contraction. Otherwise, given two flows $\dot{X}_t = -\nabla f(X_t)$ and $\dot{Y}_t = -\nabla f(Y_t)$, starting at points X_0, Y_0,*

$$\|X_t - Y_t\|^2 \leq e^{-2\alpha t} \|X_0 - Y_0\|^2$$

Corollary 12.1.2. *Given gradient flow for a differentiable and α-strongly convex f, $\dot{X}_t = -\nabla f(X_t)$, our gradient flow exponentially converges to the minimizer of our objective. This is seen by applying Theorem 12.1.1 to X_t and and using the minima of f, x^*, as Y_t.*

Proof. We have through strong-convexity,

$$\begin{aligned}\frac{d}{dt}\left\|X_t - Y_t\right\|^2 &= 2\langle X_t - Y_t, \dot{X}_t - \dot{Y}_t\rangle \\ &= -2\langle X_t - Y_t, \nabla X_t - \nabla Y_t\rangle \\ &= -2\alpha \left\|X_t - Y_t\right\|^2\end{aligned}$$

Now, by change of variables, state $U_t = \left\|X_t - Y_t\right\|^2 \geq 0$. This implies,

$$\begin{aligned}\dot{U}_t &\leq -2\alpha U_t \\ \frac{dU_t}{dt} &\leq -2\alpha U_t \\ \frac{dU_t}{U_t} &\leq -2\alpha dt \\ \int \frac{dU_t}{U_t} &\leq -2\alpha \int dt \\ \log U_t - \log U_0 &\leq -2\alpha t \\ U_t &\leq U_0 \exp[-2\alpha t] \\ \left\|X_t - Y_t\right\|^2 &\leq \exp[-2\alpha t] \left\|X_0 - Y_0\right\|^2 \square\end{aligned}$$

(This routine is known as the Grönwall inequality).

Now we would like to show a weaker result for when we have only that f is α-gradient dominated:

Theorem 12.1.3. *If f is differentiable and α-gradient dominated, along the gradient flow $\dot{X}_t = -\nabla f(X_t)$, we have exponential contraction to $x^* = \min f$.*

$$f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*))$$

Corollary 12.1.4. *By the definition of α-gradient dominated, we still have a desired result, as $\frac{\alpha}{2}\left\|X_t - x^*\right\|^2 \leq f(X_t) - f(x^*)$.*

Proof. By the gradient dominated property we have,

$$\begin{aligned}\frac{d}{dt}(f(X_t) - f(x^*)) &= \langle \nabla f(X_t), \dot{X}_t\rangle \\ &= -\left\|\nabla f(X_t)\right\|^2 \\ &= -2\alpha(f(X_t) - f(x^*))\end{aligned}$$

Again, by employing the Grönwall inequality, we have,

$$f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*)) \qquad \square$$

12.2. Optimization through discrete time

In order to discretize this process we've defined we will define a step-size $\eta > 0$, and look at 2 methods, given an arbitrary flow, $\dot{X}_t = \phi(X_t)$.

1. **Forward Method**

$$x_{k+1} = x_k + \eta\phi(x_k)$$

   Which is the same as the finite difference equation,

$$\frac{1}{\eta}(x_{k+1} - x_k) = \phi(x_k)$$

2. **Backwards Method**

$$x_{k+1} = x_k + \eta\phi(x_{k+1})$$

   Which is the same as the finite difference equation,

$$\frac{1}{\eta}(x_{k+1} - x_k) = \phi(x_{k+1})$$

   The main difference here is we are approximating the next step based on the gradient of that next step.

For gradient flow, where $\phi(x) = -\nabla f(x)$, from these two methods we get,

1. **Gradient Descent** *(forward)*

$$x_{k+1} = x_k - \eta\nabla f(x_k)$$

2. **Proimal point method** *(backward)*

$$x_{k+1} = x_k - \eta\nabla f(x_{k+1})$$

12.3. Gradient descent

We can re-express the gradient descent step as follows to obtain some properties:

$$\begin{aligned} x_{k+1} &= x_k - \eta\nabla f(x_k) \\ &= \arg\min_{x\in\mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k),\, x - x_k\rangle + \frac{1}{2\eta}\left\|x - x_k\right\|^2 \right\} \end{aligned}$$

We can confirm this is indeed the case by minimizing. The function in the $\arg\min$ is quadratic in terms of x, so we can re-write the minimization as follows:

$$
\begin{aligned}
\arg\min_{x\in\mathbb{R}^n} &\left\{ f(x_k) + \langle \nabla f(x_k),\, x - x_k\rangle + \frac{1}{2\eta}\left\|x - x_k\right\|^2 \right\} = \\
&= \arg\min_{x\in\mathbb{R}^n} \left\{ f(x_k) - \langle f(x_k),\, x_k\rangle + \langle \nabla f(x_k),\, x\rangle + \frac{1}{2\eta}\langle x - x_k,\, x - x_k\rangle \right\} \\
&= \arg\min_{x\in\mathbb{R}^n} \left\{ f(x_k) - \langle f(x_k),\, x_k\rangle + \langle \nabla f(x_k),\, x\rangle + \frac{1}{2\eta}\langle x, x\rangle - \frac{1}{\eta}\left\langle x,\, x_k\right\rangle + \frac{1}{2\eta}\left\|x_k\right\|^2 \right\} \\
&= \arg\min_{x\in\mathbb{R}^n} \left\{ \left(f(x_k) - \langle f(x_k),\, x_k\rangle + \frac{1}{2\eta}\left\|x_k\right\|^2 \right) + \left\langle \nabla f(x_k) - \frac{1}{\eta}x_k,\, x \right\rangle + \frac{1}{2}x^T\left(\frac{1}{\eta}I\right)x \right\} \\
&= \arg\min_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla f(x_k) - \frac{1}{\eta}x_k,\, x \right\rangle + \frac{1}{2}x^T\left(\frac{1}{\eta}I\right)x \right\} \\
&= \left(\frac{1}{\eta}I\right)^{-1}\left(\nabla f(x_k) - \frac{1}{\eta}x_k\right) \\
&= \eta\left(\nabla f(x_k) - \eta^{-1}x_k\right) \\
&= x_k - \eta\nabla f(x_k)
\end{aligned}
$$

In this way, we've expressed gradient descent as a **first-order** and **greedy** method. It is first order, because the first 2 terms in the $\arg\min$ corresponds to a Taylor expansion of $f(x)$ centered at x_k. However, we add the third term to regularize by how far we end up moving. This ensures that we choose an x such that we minimize the Taylor expansion, while simultaneously making sure that our Taylor expansion is still valid at the point we end up moving to. This makes it a descent method, as well, if our step=size is small enough (i.e. the regularization term is large enough to cause the Taylor expansion to stay valid, and thus we descent monotonically).

12.4. Smoothness

Recall f is L-smooth if $\forall x, y$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\left\|y - x\right\|^2$$

Which can again be re-expressed as,

$$(\nabla f(y) - \nabla f(x))^T(y - x) \leq L\left\|y - x\right\|^2$$

or,

$$\nabla^2 f(x) \preceq LI$$

We will also define the condition number of f which is α-strongly convex, and L-smooth, as $\kappa := \frac{L}{\alpha}$

12.4.1. Descent Property

We can formalize the descent property from this.

Lemma 12.4.1. *If f is L-smooth and $\eta \leq \frac{2}{L}$, then along gradient descent,*

$$f(x_{k+1}) \leq f(x_k) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(x_k)\|^2$$

- *In particular, $\eta = \frac{1}{L}$ implies,*

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\left\|\nabla^2 f(x_k)\right\|^2$$

Remark 12.4.2. We do not require convexity here.

Proof. We have, by L-smoothness:

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L}{2}\eta^2\|\nabla f(x_k)\|^2 \\
&= f(x_k) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(x_k)\|^2
\end{aligned}$$

□

12.5. Proximal Method

Now, let us analyze the Proximal method similarly. The proximal method can be written out as,

$$\begin{aligned}
x_{k+1} &= x_k - \eta\nabla f(x_k) \\
&= \arg\min_{x\in\mathbb{R}^n}\left\{f(x) + \frac{1}{2\eta}\|x - x_k\|^2\right\}
\end{aligned}$$

We have a difficulty here, as this method is implicit. To run one step, we must already minimize $f(x)$.

12.5.1. Descent Property

For the proximal method, we have instead the following lemma:

Lemma 12.5.1. *For any $\eta > 0$, along the proximal method, $f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2}\|\nabla f(x_{k+1})\|^2$.*

Remark 12.5.2. Note, this does not require convexity or smoothness.

Proof. Since x_{k+1} minimizes $f(x) + \frac{1}{2\eta}\|x - x_k\|^2$,

$$f(x_{k+1}) + \frac{1}{2\eta}\|x_{k+1} - xk\|^2 \leq f(x_k)$$

Since we have $x_{k+1} - x_k = -\eta\nabla f(x_{k+1})$,

$$
\begin{aligned}
f(x_{k+1}) &\le f(x_k) - \frac{1}{2\eta} \|x_{k+1} - x_k\|^2 \\
&= f(x_k) - \frac{\eta}{2} \|\nabla f(x_{k+1})\|^2
\end{aligned}
$$

□

12.6. Convergence rate of gradient descent

Now, given gradient descent (discrete), we would like to confirm convergence rate. We will analyze this both over convexity and weaker gradient domination.

12.6.1. Convergence rate under strong convexity

Theorem 12.6.1. *Assume f is α-strongly convex, and L=smooth, $\kappa = \frac{L}{\alpha}$. If $0\eta \le \frac{2}{\alpha+L}$, then gradient descent has exponential contraction:*

$$
\|x_k - y_k\|^2 \le \left(1 - \eta\frac{2\alpha L}{\alpha + L} \|x_0 - y_0\|^2\right)
$$

Specifically, we want when $\eta = \frac{2}{\alpha+L}$ and $y_k = x^*$,

$$
\|x_k - x^*\|^2 \le \left(1 - \frac{2}{1+\kappa}\right)^{2k} \|x_0 - x^*\|^2
$$

Proof. We will use the following lemma detailed in the work [**?**]:

Lemma 12.6.2. *If f is α-strongly convex and L-smooth, then $\forall x, y \in \mathbb{R}^d$,*

$$
\langle \nabla f(x) - \nabla f(y), \, x - y\rangle \ge \frac{\alpha L}{\alpha + L} \|x - y\|^2 + \frac{1}{\alpha + L} \|\nabla f(x) - \nabla f(y)\|^2
$$

By the descent property of gradient descent, we have:

$$
f(x_{k+1}) \le f(x_k) - \frac{1}{2\eta} \|x_{k+1} - x_k\|^2
$$

By the gradient domination property,

$$
\begin{aligned}
f(x_{k+1}) - \min f &\le f(x_k) - \min f - 2\eta\alpha \left(1 - \frac{\eta L}{2}\right) (f(x_k) - \min f) \\
&= \left(1 - 2\eta\alpha\left(1 - \frac{\eta L}{2}\right)\right) (f(x_k) - \min f)
\end{aligned}
$$

□

12.6.2. Convergence rate under Gradient Domination

We will use a similar approach.

12.7. Convergence rate of proximal method

Theorem 12.7.1. *Assume f is α-gradient dominated. For any $\eta > 0$, along proximal method, we have:*

$$f(x_k) - \min f \leq \frac{1}{(1+\eta\alpha)^k}(f(x_0) - \min f)$$

Proof. By the descent property of the proximal method,

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2} \|\nabla f(x_{k+1})\|^2$$

By gradient domination, we have,

$$f(x_{k+1}) - \min f \leq f(x_i) - \min f - \eta\alpha(f(x_{k+1}) - \min f)$$

Thus,

$$f(x_{k+1}) - \min f \leq \frac{1}{1+\eta\alpha} (f(x_k) - \min f)$$

Giving us exponential contraction. □

- Hairer, Lubich & Wanner,Geometric numerical integration:Structure-preserving algorithms for ordinary differential equations,Springer, 2006
- Karimi, Nutini & Schmidt,Linear Convergence of Gradient andProximal-Gradient Methods Under the Polyak-Lojasiewicz Condition,ECML, 2016
- Beck & Teboulle,A Fast Iterative Shrinkage-Thresholding Algorithm forLinear Inverse Problems, SIAM Journal of Imaging Sciences, 200935 References
- Boyd and Vandenberghe,Convex Optimization,CambridgeUniversityPress, 2004, available at https://web.stanford.edu/ boyd/cvxbook
- Nesterov,Lectures on Convex Optimization,Springer,2004
- Nocedal and Wright,Numerical Optimization,Springer,2006
- Vishnoi, Algorithms for Convex Optimization, available at https://convex-optimization.github.ioCPSC 463:Algorithms for Continuous Optimization,Fall2021
- Recht, Optimization, Big Data Bootcamp at Simons Institute, 2013, https://simons.berkeley.edu/talks/ben-recht-2013-09-04
- Boyd,Convex Optimization, Lecture videos for EE 364A at Stanford, 2008, https://www.youtube.com/watch?v=McLq1hEq3UY