10-425/625: Introduction to Convex Optimization **(Fall 2023)**

Lecture 8: Convergence of Gradient Descent

Instructor:[1] *Matt Gormley* *September 20, 2023*

8.1 Gradient Descent

Recall from a previous lecture

Recall the gradient descent algorithm:

- Choose initial point $x^0 \in \mathbb{R}^n$
- Repeat:
$$x^{t+1} = x^t - \eta_t \nabla f(x^t), \quad t = 1, 2, 3, \ldots$$
- Stop when $\|\nabla f(x^t)\|_2^2$ is small

8.2 Two Canonical Examples

It is worth studying gradient descent in two simple analytical examples to understand the type of behavior we might expect.

8.2.1 Problem 1: Least Squares

Suppose we are solving a least squares problem:

$$\min \frac{1}{2}\|Ax - b\|_2^2,$$

where $S := A^T A$ has finite condition number, i.e.

$$\kappa(S) = \frac{\lambda_{\max}(S)}{\lambda_{\min}(S)} < \infty.$$

[1] These notes were originally written by Siva Balakrishnan for 10-725 Spring 2023 (original version: here) and were edited and adapted for 10-425/625.

This is equivalent to saying our problem is both smooth and strongly convex (the most favorable case for GD).

Here we know the solution in closed form:

$$\widehat{x} = (A^T A)^{-1} A^T b,$$

and in particular we can write $\widehat{x}$ as the (only) solution to the linear system $(A^T A)\widehat{x} = A^T b$—this system of equations is called the *normal equations*. However, we might wish to avoid computing and inverting the covariance matrix, and instead simply use GD on the least squares objective.

Now, observe that the gradient of the objective, is $\nabla f(x) = -A^T(b - Ax)$ so that, the gradient descent iteration is simply,

$$\begin{aligned} x^{t+1} &= x^t + \eta A^T(b - Ax^t) \\ &= x^t - \eta A^T(Ax^t - b) \\ &= x^t - \eta A^T A x^t + \eta A^T b. \end{aligned}$$

Subtracting $\widehat{x}$ from both sides, and substituting the left side of the normal equations for $A^T b$ from above, then rearranging, we can see that,

$$\begin{aligned} x^{t+1} - \widehat{x} &= x^t - \widehat{x} - \eta A^T A x^t + \eta A^T b \\ &= x^t - \widehat{x} - \eta A^T A x^t + \eta (A^T A)\widehat{x} \\ &= \left[I - \eta(A^T A)\right](x^t - \widehat{x}). \end{aligned}$$

We can unroll this to see that after k time steps x^k satisfies,

$$x^k - \widehat{x} = \left[I - \eta(A^T A)\right]^k (x^0 - \widehat{x}),$$

as a direct consequence we see that,

$$\|x^k - \widehat{x}\|_2 \leq \|I - \eta(A^T A)\|_{\text{op}}^k \|x^0 - \widehat{x}\|_2.$$

So if we can ensure that the operator norm term < 1 we will have rapid (geometric) decay of the distance between our iterate and the optimal solution.

Background: (Operator Norm) For a square matrix $A \in \mathbb{R}^{n \times n}$, the *operator norm* is given by:

$$\|A\|_{\text{op}} = \inf\{c \geq 0 : \|Ax\|_2 \leq c\|x\|_2, \forall x \in \mathbb{R}^n\}$$

For any square matrix A, the operator norm $\|A\|_{\text{op}}$ is equal to the largest singular value of the matrix A.

For any symmetric matrix A, the operator norm $\|A\|_{\text{op}}$ is equal to the largest eigenvalue (since, for any symmetric matrix, the largest singular value equals the largest eigenvalue).

Some useful properties of the operator norm:

- $\|Ax\|_2 \leq \|A\|_{\text{op}}\|x\|_2, \forall x \in \mathbb{R}^n$
- $\|cA\|_{\text{op}} = |c|\|A\|_{\text{op}}, \forall c \in \mathbb{R}$
- $\|A + B\|_{\text{op}} \leq \|A\|_{\text{op}} + \|B\|_{\text{op}}$, for square matrices A, B
- $\|AB\|_{\text{op}} \leq \|A\|_{\text{op}}\|B\|_{\text{op}}$, for square matrices A, B

Let us denote $A^T A := S$. Now, one can check the following fact: if we choose $\eta = \frac{2}{\lambda_{\max}(S)+\lambda_{\min}(S)}$ (this is some ideal choice that we won't have access to in practice, but will help us in theory) then $\|I - \eta(A^T A)\|_{\text{op}} = (\lambda_{\max}(S) - \lambda_{\min}(S))/(\lambda_{\max}(S) + \lambda_{\min}(S)) = (\kappa(S) - 1)/(\kappa(S) + 1) < 1$ and we see that,

$$\|x^k - \widehat{x}\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x^0 - \widehat{x}\|_2.$$

Some notes about this result:

1. We might sometimes (often) care instead about the value of the objective function at our iterates i.e. we would like to upper bound $f(x^k) - f(\widehat{x})$. For nice quadratics its easy to obtain a bound on this

error from a bound on $\|x^k - \widehat{x}\|_2$. Some algebra will show that,

$$\begin{aligned} f(x^k) - f(\widehat{x}) &= \frac{(x^k - \widehat{x})^T A^T A (x^k - \widehat{x})}{2} \\ &\leq \frac{\lambda_{\max}(S)}{2} \|x^k - \widehat{x}\|_2^2 \\ &\leq \frac{\lambda_{\max}(S)}{2} \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - \widehat{x}\|_2^2. \end{aligned}$$

2. This type of convergence is often called *linear* convergence in optimization (and sometimes called geometric convergence). A consequence of the above statement is that if I want my error to be $\leq \epsilon$ then it suffices to take $k \sim \log(1/\epsilon)$ steps (ignoring constants which depend on how far you initialize, and the condition number of S).

8.2.2 Problem 2: Univariate Absolute Value

Another prototypical example is applying (sub)GD to the (univariate) function $f(x) = |x|$. Suppose that we initialize at some point $x_0 = -1$, and use some constant step-size $\eta = 0.7$ (ignoring for now the non-differentiability at 0). We can see that in general the GD iterates will bounce around the optimum, and will not converge. The iterates will be $x_t = -1, -0.3, -0.4, -0.3$. A picture is easier to follow.

In this case, the only way to "force" GD to converge will be to use a decaying step-size (or if we want to get to within ϵ of the optimum we should use a step-size that is smaller than that), and this will result in much slower convergence.

This is one of the main problems of trying to optimize functions which are not smooth.

8.3 GD Convergence Results

For the rest of this lecture, we'll assume that our objective function f is twice-differentiable and β-smooth. Our goal will be to try to understand the behaviour of GD in three settings which are increasingly "nicer":

1. Arbitrary (possibly non-convex) function f which is twice-differentiable and β-smooth.
2. Convex function f which is twice-differentiable and β-smooth.
3. Convex function f which is twice-differentiable and β-smooth, and is additionally α-strongly convex.

Most of these results don't require twice-differentiability but the proofs are sometimes a bit more transparent when you do have twice-differentiability.

Definition 8.1 (ϵ-suboptimal)**.** *For an optimization problem with minimizer x^*, a point x is* ϵ-suboptimal *if*

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

Definition 8.2 (ϵ-substationary)**.** *A point x is* ϵ-substationary *if*

$$\|\nabla f(x)\|_2 \leq \epsilon$$

This could occur near a saddle point, a local minimum, or a local maximum where the gradient is exactly zero.

8.3.1 Analysis for smooth, (possibly) nonconvex case

Assume f is differentiable, possibly nonconvex, and β-smooth. Recall that the latter means that $\nabla f(x)$ is Lipschitz with constant $\beta > 0$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2 \quad \text{for any } x, y$$

Or, equivalently, when twice differentiable: $\nabla^2 f(x) \preceq LI$.

Under these assumptions, asking for ϵ-suboptimality is too much. Let's settle for a ϵ-substationary point x, which means $\|\nabla f(x)\|_2 \leq \epsilon$.

Theorem 8.3. *Gradient descent with fixed step size $\eta \leq 1/\beta$ satisfies*

$$\min_{t=0,\ldots,k} \|\nabla f(x^{(t)})\|_2 \leq \sqrt{\frac{2\beta}{\eta}(f(x^{(0)}) - f(x^*))}$$

Thus gradient descent has rate $O(1/\sqrt{k})$, or $O(1/\epsilon^2)$, even in the nonconvex case for finding stationary points.

This rate *cannot be improved* (over class of differentiable functions with Lipschitz gradients) by any deterministic algorithm.

8.3.2 Analysis for smooth, convex case

Assume that f convex and differentiable, with $\text{dom}(f) = \mathbb{R}^n$, and additionally that f is β-smooth.

Theorem 8.4. *Gradient descent with fixed step size $\eta \leq 1/\beta$ satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\beta}{2k}\|x^{(0)} - x^\star\|_2^2$$

We say gradient descent has convergence rate $O(1/k)$. That is, it finds ϵ-suboptimal point in $O(1/\epsilon)$ iterations.

8.3.3 Analysis for smooth, strongly convex case

Reminder: *strong convexity* of f means $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex for some $\alpha > 0$ (when twice differentiable: $\nabla^2 f(x) \succeq \alpha I$).

Assuming Lipschitz gradient as before, and also strong convexity:

Theorem 8.5. *Gradient descent with fixed step size $\eta \leq 2/(\alpha + \beta)$ or with backtracking line search search satisfies*

$$f(x^{(k)}) - f(x^*) \leq \gamma^k \frac{\beta}{2}\|x^{(0)} - x^\star\|_2^2$$

where $0 < \gamma < 1$

Rate under strong convexity is $O(\gamma^k)$, exponentially fast! That is, it finds ϵ-suboptimal point in $O(\log(1/\epsilon))$ iterations.