Lecture 4 : General Descent Methods

EE227C . Lecturer: Professor Martin Wainwright . Scribe: Alvin Wan

Recall $Q \leq R \leftrightarrow R - Q \geq 0$ or ($R - Q$ is positive semidefinite.)

1 Strongly Convex Gradient Descent

Theorem: Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, m-strongly convex, and m-smooth. Then, gradient descent $x^{l+1} = x^l - \alpha \nabla f(x^l)$ satisfies

$$\|x^l - x^*\|_2 \leq (1 - \frac{m}{M})^l \|x^0 - x^*\|_2 \forall l = 1, 2, 3 \ldots$$

where $\alpha = \frac{2}{M+m}$ and $\frac{m}{M}$ is the inverse condition number.

The above improves on our result from the previous note. Instead of considering second derivatives, we can consider an equivalent definition.

$$\frac{m}{2}\|x - y\|_2^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{M}{2}\|x - y\|_2^2 \forall x, y \in \mathbb{R}^d$$

Start from x and extrapolate to y. Our linear extrapolation $f(x) + \langle \nabla f(x), y - x \rangle$. Smoothness means that it grows no faster than a quadratic with coefficient $\frac{M}{2}$ in front. Convexity means it grows no slower than a quadratic with $\frac{m}{2}$ in front. This in essence, is what it means to sandwich in between two quadratics. Let us call the function that satisfies these bounds $(m, M) - f$.

Lemma: Any $(m, M) - f$ satisfies:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{mM}{m + M}\|x - y\|_2^2 + \frac{1}{M + m}\|\nabla f(x) - \nabla f(y)\|_2^2$$

This lemma says the following: we can get a lower bound on the difference in gradients and points. We found last time that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \|x - y\|_2^2$. What we'll find this time is a lower bound with coefficients in terms of m, M. To prove the lemma, consider $\phi(x) = f(x) - \frac{m}{2}\|_2^2$, which is convex and $(M - m) - smooth$. Let us begin by proving the theorem using the lemma as given. Last time, we did the following noting that $x^{l+1} - x^l = -\alpha(\nabla f(x^l) - \nabla f(x^*))$:

$$
\begin{aligned}
&\|x^{l+1} - x^*\|_2^2 \\
&= \|x^{l+1} - x^l + x^l - x^*\|_2^2 \\
&= \|x^l - x^*\|_2^2 + \alpha^2 \|\nabla f(x^l) - \nabla f(x^*)\|_2^2 - 2\alpha \langle \nabla f(x^l) - \nabla f(x^*), x^l - x^* \rangle \\
&\leq (1 - 2\alpha \frac{mM}{m+M}) \|x^l - x^*\|_2^2 + \alpha(\alpha - \frac{2\alpha}{m+M}) \|\nabla f(x^l) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

Plugging in $\alpha = \frac{2}{m+M}$, the last term drops off. Hence,

$$
\begin{aligned}
\|x^{l+1} - x^*\|_2^2 \leq (1 - \frac{4}{m+M} \frac{mM}{m+M} \|x^l - x^*|_2^2 \\
= (\frac{m-M}{m+M})^2 \|x^l - x^*\|_2^2
\end{aligned}
$$

This implies the following, where $\kappa = \frac{M}{m}$.

$$
\|x^l - x^*\|_2 \leq (\frac{1 - 1/\kappa}{1 + 1/\kappa})^l \|x^l - x^*\|_2
$$

This is stronger than the lemma we set out to prove. Before, we directly bounded the gradient and eventually had a complicated term in terms of α. This time, we leveraged both the smoothness and convexity in a more efficient way.

How many steps do I need so that $\|x^N - x^*\|_2 \leq \epsilon \leftarrow (\frac{1-1/\kappa}{1+1/\kappa})^N \|x^0 - x^*\|_2 \leq \epsilon$. We have $N(\epsilon) = \log(\frac{\|x^0 - x^*\|_2}{\epsilon})(\log(\frac{1+1/\kappa}{1-1/\kappa}))^{-1}$.

2 Contractivity

As we noted last time, this bound isn't practical to compute, because computing M, m is expensive to compute. As a result, we will develop more sophisticated notions to choose step sizes.

We showed the operator $T_\alpha : \mathbb{R}^d \to \mathbb{R}^d$ is the following:

$$
T_\alpha(x) = x - \alpha \nabla f(x) \text{with } \alpha = \frac{2}{m+M}
$$

contractive with parameter $\gamma = \frac{1-1/\kappa}{1+1/\kappa}$. The following gives us gradient descent with errors.

$$x^{l+1} = T(x^l) + e^l$$

It can be hard to evaluate the function value let alone compute its gradients. It could also be that gradient computation is expensive or inaccurate. Why contractivity is important: If you have a contraction, it will a) be stable under determinstic errors with $l-2$ norm. In other words, your algorithm doesn't suddenly diverge; it will converge to a neighborhood of x^*. Even with random noise that is bounded in l-2 norm, it will converge to a neighborhood of x^*. Long story short: when you can prove iterates are contractive, you can show that the algorithm is resistant to noise.

3 Convex Gradient Descent

We might want to deal with functions that are not differentiable. For example, take the l-1 norm. You can still have functions that are not differentiable but still convex, such as $|x|$. These arise in practice, so we will need to handle these types of functions. We additionally need to handle more misbehaving functions. To do this, we need to relax conditions and see what else we can guarantee. We pay a price, however, with looser constraints. In some cases, we don't even have a function that converges. We will need to modify constraints incrementally to see how it behaves.

3.1 Descent Methods

We now ask: what happens if we have less structure? We now remove the strong convexity condition, setting $m = 0$. Suppose f is convex, differentiable, and M-smooth.

$$x^{l+1} = x^l - \frac{1}{M}\nabla f(x^l) = \text{argmin}_{y\in\mathbb{R}^d}\{f(x^l) + \langle \nabla f(x^l), y - x^l\rangle + \frac{M}{2}\|y - x^l\|_2^2\}$$

Note that this is gradient descent with step size $\alpha = \frac{1}{M}$. The second expression is a linear approximation at x^l summed with a quadratic regularizer. This tells us that gradient method with $\alpha = \frac{1}{M}$ is a **descent method**, meaning a method that generates iterates such that the error is decreasing.

$$f(x^{l+1}) - f(x^l) - \langle \nabla f(x^l), x^{l+1} - x^l \rangle \leq \frac{M}{2} \|x^{l+1} - x^l\|_2^2$$

Note $x^{l+1} - x^l = -\frac{1}{M} \nabla f(x^l)$, $\|x^{l+1} - x^l\|_2^2 = \| - \frac{1}{M} \nabla f(x^l)\|_2^2$, and $\langle \nabla f(x^l), x^{l+1} - x^l \rangle = -\frac{1}{M} \|\nabla f(x^l)\|_2^2$. We then get taht

$$f(x^{l+1}) \leq f(x^l) - \frac{1}{2M} \|\nabla f(x^l)\|_2^2$$

This means our method is indeed a descent method. Suppose $f(x^*) = \min_{x \in \mathbb{R}^d} f(x)$ exists and is finite. We don't have guarantees for the uniqueness of x^*, so we can't discuss convergence to a particular x^*. We can consider converge of function values $f(x^l) \to f(x^*)$ or convergence to this set. We will choose the former, but the rate will not be especially fast.

The more general class of methods picks a direction that decreases cost. The newton method, for example, is much more efficient in some cases. We will also want to understand coordinate descent. Instead of computing the full gradient, we pick one gradient to compute a gradient for. We then step in the direction of that direction. We'd expect coordinate descent to perform at a slower rate, but it's $O(d)$ faster to compute. The name "stochastic gradient" is a misnomers, as they are not descent methods and do not decrease a cost function.

3.2 Rate of Convergence

Theorem Say f is convex, differentiable, and M-smooth. Say $f(x^*) = \min_{x \in \mathbb{R}^d} f(x), x^* \in \text{argmin}_{x \in \mathbb{R}^d} f(x)$. Then, gradient descent with $\alpha = \frac{1}{M}$ step size:

$$f(x^T) - f(x^*) \leq \frac{M \|x^0 - x^*\|_2^2}{2T}$$

In other words, after T steps, we will reach an error that drops to a constant over T.This is much slower than the previous rate, exponentially slower in fact. The iteration complexity of this algorithm is the following.

$$N(\epsilon) = \frac{M \|x^0 - x^*\|_2}{2\epsilon}$$

Contrast this with $\log(\frac{1}{\epsilon})$ with strong convexity. We're only toying with polynomial factors, without the guarantee of convexity. We have that

$$
\begin{aligned}
& f(x^{l+1}) \leq f(x^l) \\
& = \frac{1}{2M}\|\nabla f(x^l)\|_2^2 \leq f(x^*) + \langle \nabla f(x^l), x^* - x^l \rangle - \frac{1}{2M}\|\nabla f(x^l)\|_2^2 \\
& = f(x^*) + \frac{M}{2}\{\|x^l - x^*\|_2^2 = \|x^l - x^* - \frac{1}{M}\nabla f(x^l)\|_2^2\}
\end{aligned}
$$

Now, we have a telescoping recursion in the last term, when we plug in. The sequence is decreasing, so by average, we get a smaller value. This first step is due to the fact that $\{f(x^l) - f(x^*)\}_{l=0}^{\infty}$ is decreasing. In the last bound, we ignore the negative term.

$$
f(x^T) - f(x^*) \leq \frac{1}{T}\sum_{l=0}^{T-1}(f(x^{l+1}) - f(x^*)) \leq \frac{M}{2T}\{\|x^0 - x^*\|_2^2\}
$$

The proven bound works for all minimizers, as we make no additional assumptions about x^*. We can thus replace it with the infimum over all such minimizers.