Convergence of iterates of gradient descent

Sunday, March 28, 2021 2:03 PM

We proved that if f is convex and ∇f exists and is L-Lipschitz cts, then gradient descent $(x_k)_{k=0}^{\infty}$, $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$, converges in the sense that $f(x_k) \to f^* := \min_x f(x)$. (we also have a rate)

Q: does (x_k) itself converge? It's possible for $f(x_k)$ to converge but (x_k) to not (ex: $f(x) := 0$, $x_k = (-1)^k$). Can this happen w/ gradient descent?

Easy case: f is strongly convex. Then $\left(f(x_k) \to f^*\right) \Rightarrow \left(x_k \to x^*\right)$ ... and we have a rate.

and $x^* = \operatorname{argmin} f(x)$ is unique.

This note: we'll prove that even if f isn't strongly convex, the sequence generated by gradient descent still converges. (Assuming f convex, ∇f Lipschitz, and $x_k \in \mathbb{R}^n$. Our argument fails in ∞-dim.)
(but we won't have a rate of convergence)

proof

Recall in our proof for convergence of $f(x_k)$ for gradient descent, using the descent lemma and convexity, we proved:

$$f(x_k) - f^* \le \underbrace{\|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2}$$

before, we then summed and noticed this telescoped.

Since $f(x_k) \ge f^*$ this means

any optimal point (since may not be unique)

$$\|x_k - x^*\| \le \|x_{k-1} - x^*\| \quad \forall\, k = 1, 2, \ldots$$

Such a sequence is called Fejér monotone

Leopold Fejér, pr. "Fay-er", Hungarian 1922

which immediately implies $(\|x_k - x^*\|)_k$ is bounded, hence $(\|x_k\|)_k$ is bounded.

*Note: not true in infinite dimensions

i.e., (x_k) is contained in a compact set*, hence it has at least one convergent subsequence. We'll use:

Proposition If (x_k) is in a compact set in a Banach space, then x_k converges iff *all* convergent subsequences have the same limit

(proof sketch in $\mathbb{R}$: pick subseq converging to $\limsup x_k$ and $\liminf x_k$. In general this proof doesn't extend but there are other proofs)

So, let $x_{k_j} \to x$ be one convergent subseq., and $x_{n_j} \to y$ another.

Goal is to show $x = y$ and that x is optimal.

Easy: both x and y optimal since ∇f exists $\Rightarrow$ f is (sequentially) continuous,

so $x_{k_j} \to x \Rightarrow f(x_{k_j}) \to f(x)$

but $f(x_k) \to f^*$ so $f(x_{k_j}) \to f^*$ also

hence $f(x) = f^*$ meaning it's optimal.

... which applies since we proved x is optimal

Now by Fejér monotonicity,

$\|x_k - x\|$ is bounded and non-increasing, so it's convergent, with some limit $\alpha = \lim_k \|x_k - x\|$

Thus $\|x_{k_j} - x\| \to \alpha$ as well since it's a subsequence,

but since we've assumed $x_{k_j} \to x$ this means $\alpha = 0$

But now we're done: two ways to see it

1) $\|x_{n_j} - x\| \to \alpha = 0$ as well, so by sequential continuity of norms, $\|y - x\| = 0$ meaning $y = x$.

Argument via 2011 Beck, Teboulle chapter

2) or, since $\|x_k - x\| \to \alpha = 0$ this is definition of convergence, so $x_k \to x$ as desired (in particular, all subseq. have same limit so $y = x$).

In fact w/ this 2nd approach, we didn't even need to use the proposition!

Q: Is the same true for Nesterov accelerated gradient descent?

A: sort of. For most variants, there are no published proofs that it does,* though I've heard "through-the-grapevine" that it is possible to show it.

*as of ≈2020

But there are definitely some particular accelerated variants that do have guaranteed convergence of (x_k),

Ex: "On the convergence of the iterates of the 'FISTA' algorithm" Chambolle and Dossal, 2015, J. Optim. Theory + Applic.

Does this kind of stuff interest you?

Learn more (e.g., what if $x_k \in \mathcal{H}$ for some ∞-dimensional Hilbert space)

in Bauschke + Combettes' book

"Convex Analysis + Monotone Operator Theory in Hilbert Spaces", 2nd ed. 2017