Unified analysis of gradient/subgradient descent
APPM 5630 Advanced Convex Optimization

Instructor: Prof. Becker **Revision date**: 1/29/24

We'll solve $\min_{\mathbf{x}} \ f(\mathbf{x})$ via the following generic algorithm, with $R = \|\mathbf{x}_1 - \mathbf{x}^\star\|$,

---

Require: $\mathbf{x}_1$
1: **for** $k = 1, 2, \ldots, K$ **do**
2: $\quad \mathbf{x}_{k+1} = \mathbf{x}_k - t\mathbf{v}_k$
3: **end for**

---

where $\mathbf{v}$ is "gradient-like" (e.g., a gradient, subgradient, or a gradient in expectation, like $\mathbb{E}\,\mathbf{v}_k = \nabla f(\mathbf{x}_k)$).

Lemma 1 (Lemma 14.1 in Shalev-Shwartz and Ben-David)**.** *Let $\{\mathbf{v}_k\}_{k=1}^K$ be arbitrary. No assumptions on f (need not be convex or smooth). The generic algorithm sequence satisfies*

$$\sum_{k=1}^K \langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{v}_k \rangle \le \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2t} + \frac{t}{2}\sum_{k=1}^K \|\mathbf{v}_k\|^2 \tag{1}$$

Proof. (Sketch: just the good parts)

$$\begin{aligned}
\sum_{k=1}^K \langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{v}_k \rangle &= \frac{1}{2t}\sum_{k=1}^K \left(-\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^2 + \|\mathbf{x}_k - \mathbf{x}^\star\|^2 + t^2\|\mathbf{v}_k\|^2\right) \quad \text{complete-the-square and algebra} \\
&= \frac{1}{2t}\left(\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{K+1} - \mathbf{x}^\star\|^2\right) + \frac{1}{2t}\sum_{k=1}^K t^2\|\mathbf{v}_k\|^2 \quad \text{via telescoping sum} \\
&\le \frac{1}{2t}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \frac{t}{2}\sum_{k=1}^K \|\mathbf{v}_k\|^2.
\end{aligned}$$

□

A variant of the above result, using a possibly non-constant stepsize $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k\mathbf{v}_k$, is known as Shor's Hyperplane Distance Convergence from his 1985 book; see Theorem 1 in Convergence Rates for Deterministic and Stochastic Subgradient Methods Without Lipschitz Continuity by Benjamin Grimmer (2019 SIOPT) for an elementary proof.

Lemma 2 (Shor's Hyperplane Distance lemma)**.**

$$\langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{v}_k/\|\mathbf{v}_k\| \rangle \le \frac{\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \sum_{k=1}^K t_k^2\|\mathbf{v}_k\|^2}{2\sum_{k=1}^K t_k\|\mathbf{v}_k\|}$$

Going back to the simpler Lemma 1, a basic corollary is the following:

Corollary 3 (2nd part of Lemma 14.1)**.** *If $\|\mathbf{v}_k\| \le \rho$ (e.g., if f is ρ-Lipschitz) and $t = \frac{R}{\rho\sqrt{K}}$ then*

$$\frac{1}{K}\sum_{k=1}^K \langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{v}_k \rangle \le \rho\frac{R}{\sqrt{K}}$$

Proof. Plugging in $\|\mathbf{v}_k\|^2 \le \rho^2$ and $R^2 = \|\mathbf{x}_1 - \mathbf{x}^\star\|^2$ into the RHS of Eq. (1) gives $\frac{1}{2}\left(R^2/t + tK\rho^2\right)$ which is minimized at the given value of t leading to $\rho R\sqrt{K}$. Dividing the LHS and RHS of Eq. (1) by K gives the result. □

Now we'll see how to use these results.

1 f is convex but not smooth

Assume f is ρ-Lipschitz so the corollary applies. If f is convex, then we have a well-defined subdifferential, so we'll choose $\mathbf{v}_k \in \partial f(\mathbf{x}_k)$ to give us **subgradient descent**. By convexity and definition of subgradients,

$$f(\mathbf{x}_k) - f^\star \leq \langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{v}_k \rangle \tag{2}$$

so combining this with Corollary 3 immediately yields

Corollary 4 (sub-gradient descent, Cor. 14.2)**.** *If f is convex and ρ-Lipschitz, then subgradient descent (with $t = \frac{R}{\rho\sqrt{K}}$) yields*

$$\frac{1}{K}\sum_{k=1}^{K}\left(f(\mathbf{x}_k) - f^\star\right) \leq \rho\frac{R}{\sqrt{K}}$$

hence

$$f(\mathbf{x}_{best}) - f^\star \leq \rho\frac{R}{\sqrt{K}} \tag{3}$$

and

$$f(\bar{\mathbf{x}}) - f^\star \leq \rho\frac{R}{\sqrt{K}} \tag{4}$$

where $\mathbf{x}_{\text{best}} \in \operatorname{argmin}_{\mathbf{x}\in\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_K\}} f(\mathbf{x})$ and $\bar{\mathbf{x}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k$. If possible, we should use $\mathbf{x}_{\text{best}}$, but in some situations this is not easy. Subgradient descent is not a descent method, so it's not necessarily true that $\mathbf{x}_{\text{best}} = \mathbf{x}_K$. Couldn't we just evaluate $f(\mathbf{x}_k)$ and record the best iterate seen so far? Often we can do this, but sometimes f is very expensive to evaluate (as will especially be the case when we do *stochastic* gradients which sample, and the true loss function f is a population expectation that we can never calculate). In these case, we can do iterate averaging to get $\bar{\mathbf{x}}$, and this result follows because $f(\bar{\mathbf{x}}) \leq \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}_k)$ via Jensen's inequality.

Commentary Unlike gradient descent in the smooth case, here we have slower convergence $1/\sqrt{K}$ vs $1/K$ in the smooth case (or $1/K^2$ for Nesterov acceleration). Furthermore, we need to know the maximum number of iterations K in advance in order to set the stepsize. In practice, like stochastic gradient methods, one might use a constant stepsize for a while, then reduce it: a stepsize "schedule."

2 f is smooth (∇f is L-Lipschitz continuous)

We use the descent Lemma, which applies whenever ∇f is L-Lipschitz continuous, regardless of convexity:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

and when applied to $\mathbf{x} = \mathbf{x}_k$, $\mathbf{y} = \mathbf{x}_k - t\mathbf{v}_k$ with $\mathbf{v}_k = \nabla f(\mathbf{x}_k)$ and $t = L^{-1}$ (this is **gradient descent**) after a bit of algebra gives

$$f(\mathbf{x}_{k+1}) \overset{\text{descent lem.}}{\leq} f(\mathbf{x}_k) + \left(\frac{L}{2}t^2 - t\right)\|\nabla f(\mathbf{x}_k)\|^2 \overset{t=L^{-1}}{\leq} f(\mathbf{x}_k) - \frac{1}{2L}\|\underbrace{\nabla f(\mathbf{x}_k)}_{\mathbf{v}_k}\|^2. \tag{5}$$

Also using that $f(\mathbf{x}^\star) \leq f(\mathbf{x}_{k+1})$ gives us another useful result:

$$f(\mathbf{x}) - f^\star \geq \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2. \tag{6}$$

If we don't assume f is convex, we can't expect to converge to the global minimizer, so there isn't a result about $f(\mathbf{x}_k) - f^\star \to 0$. Instead, we show convergence to a stationary point, meaning $\|\nabla f(\mathbf{x}_k)\| \to 0$.

Corollary 5 (gradient descent, non-convex)**.** *If ∇f L-Lipschitz, then gradient descent with $t = L^{-1}$ yields*

$$\min_{k=1,\ldots,K} \|\nabla f(\mathbf{x}_k)\|^2 \le \frac{2L}{K}(f(\mathbf{x}_1) - f^\star)$$

Proof. Sum Eq. (5) from $k = 1, \ldots, K$ after re-arranging to get

$$\frac{1}{2L}\sum_{k=1}^{K} \|\nabla f(\mathbf{x}_k)\|^2 \le \sum_{k=1}^{K} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) = f(\mathbf{x}_1) - f(\mathbf{x}_{K+1}) \le f(\mathbf{x}_1) - f^\star$$

since we had a telescoping series, and use $\min_{k=1,\ldots,K} \|\nabla f(\mathbf{x}_k)\|^2 \le \frac{1}{K}\sum_{k=1}^{K} \|\nabla f(\mathbf{x}_k)\|^2$ since the min is less than the average. ☐

In the convex case, we expect to converge to the global minimizer:

Corollary 6 (gradient descent, convex)**.** *If ∇f L-Lipschitz, and f is convex, then gradient descent with $t = L^{-1}$ yields*

$$f(\mathbf{x}_{K+1}) - f^\star \le \frac{L}{2K}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2.$$

Proof. Using the main Lemma (Eq. 1) and replacing $\langle \mathbf{x}_k - \mathbf{x}^\star, \mathbf{v}_k\rangle$ with the bound in Eq. (2) (since gradients are subgradients) gives

$$\sum_{k=1}^{K} f(\mathbf{x}_k) - f^\star \le \frac{1}{2t}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \frac{t}{2}\sum_{k=1}^{K}\|\underbrace{\nabla f(\mathbf{x}_k)}_{\mathbf{v}_k}\|^2 \tag{7}$$

and the descent lemma Eq. (5) gives $f(\mathbf{x}_{k+1}) + \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \le f(\mathbf{x}_k)$, so combining with the above equation gives

$$\begin{aligned}\sum_{k=1}^{K}\left(f(\mathbf{x}_{k+1}) + \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 - f^\star\right) &\le \sum_{k=1}^{K} f(\mathbf{x}_k) - f^\star \quad \text{via descent lemma}\\ &\le \frac{L}{2}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \frac{1}{2L}\sum_{k=1}^{K}\|\nabla f(\mathbf{x}_k)\|^2 \quad \text{via Eq. (7)}\end{aligned}$$

where we used $t = 1/L$. Now canceling the $\frac{1}{2L}\sum_{k=1}^{K}\|\nabla f(\mathbf{x}_k)\|^2$ from both sides gives

$$\sum_{k=1}^{K} f(\mathbf{x}_{k+1}) - f^\star \le \frac{L}{2}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2$$

hence

$$f(\mathbf{x}_{K+1}) = f(\mathbf{x}_{\text{best}}) \le \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}_{k+1}) - f^\star \le \frac{L}{2K}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2$$

where $\mathbf{x}_{K+1} = \mathbf{x}_{\text{best}}$ follows because the descent lemma implies that this is a descent method. ☐

A variant of the above fixed-stepsize case is to use the "Polyak" adaptive stepsize with $t_k = \frac{f(\mathbf{x}_k) - f^\star}{\|\nabla f(\mathbf{x}_k)\|^2}$ or similar (so $t_t \ge 1/(2L)$ if ∇f is L-Lipschitz). For proof techniques using that stepsize, see Revisiting the Polyak Step Size by Elad Hazan and Sham M. Kakade (2019).

Our last case to consider is if we're **strongly convex**, in which case we expect faster convergence, and $\mathbf{x}^\star$ is unique, and we expect a bound on $\|\mathbf{x}_k - \mathbf{x}^\star\|$. Note that if f is μ strongly convex, then f satisfies the μ Polyak-Lojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \ge \mu(f(\mathbf{x}) - f^\star) \tag{8}$$

(see Nesterov's 2018 book, Thm 2.1.5 and Eq 2.1.10 for a proof). Our result is

Corollary 7 (gradient descent, strongly convex)**.** *If ∇f L-Lipschitz, and f is μ strongly convex, then gradient descent with $t = L^{-1}$ yields*

$$f(\mathbf{x}_{K+1}) - f^\star \leq \underbrace{\left(1 - \frac{\mu}{L}\right)}_{c}^{K-1} (f(\mathbf{x}_1) - f^\star)\,.$$

This is linear convergence, which is asymptotically better than sublinear convergence. We think of $\kappa = \frac{L}{\mu}$ as the condition number, so $c = 1 - \kappa^{-1}$. We won't show it here, but Nesterov acceleration can improve c to $c \approx 1 - \kappa^{-1/2}$ when $\kappa \gg 1$.

Proof.

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \frac{-1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{-\mu}{L} (f(\mathbf{x}_k) - f^\star)$$

using the descent lemma for the first inequality and the PL inequality for the second inequality. Re-arranging and recursing gives

$$f(\mathbf{x}_{k+1}) - f^\star \leq \left(1 - \frac{\mu}{L}\right) (f(\mathbf{x}_k) - f^\star) \leq \left(1 - \frac{\mu}{L}\right)^{t-1} (f(\mathbf{x}_1) - f^\star)\,.$$

□