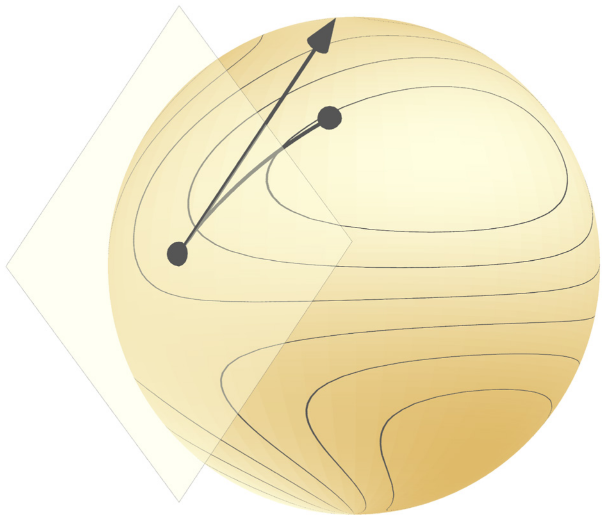
203

Gradient descent

Spring 2023

Optimization on manifolds, MATH 512 @ EPFL

Instructor: Nicolas Boumal



A family of gradient descent methods

We aim to minimize $f: \mathcal{M} \to \mathbf{R}$, smooth on a manifold.

Choose a retraction R, a Riemannian metric on $\mathcal{M}$, and $x_0 \in \mathcal{M}$.

Algorithm template for (Riemannian) gradient descent:

$\forall k = 0, 1, 2, \ldots$: $\quad x_{k+1} = R_{x_k}(-\alpha_k \operatorname{grad} f(x_k))$, for some step-size $\alpha_k > 0$.

$$f(x_{k+1}) = f(R_{x_k}(\ldots\ldots)) \ ?$$

Aim for minima, guarantee small gradients

Recall $f\big(\underbrace{R_x(s)}_{x_{k+1}}\big) = f(\underbrace{x}_{x_k}) + \langle \mathrm{grad} f(x), \underbrace{s}_{-\alpha_k \mathrm{grad} f(x_k)} \rangle_x + \underbrace{O(\|s\|_x^2)}_{O(\alpha_k^2 \|\mathrm{grad} f(x_k)\|_{x_k}^2)}$ for $s \in \mathrm{T}_x\mathcal{M}$.

$$f(x_{k+1}) = f(x_k) - \alpha_k \|\mathrm{grad} f(x_k)\|_{x_k}^2 + O(\alpha_k^2) \|\mathrm{grad} f(x_k)\|_{x_k}^2$$

$$f(x_k) - f(x_{k+1}) = \left(\alpha_k - O(\alpha_k^2) \right) \|\mathrm{grad} f(x_k)\|_{x_k}^2$$

$x_{k+1} = R_{x_k}(-\alpha\, \mathrm{grad} f(x_k))$

A simple result for constant step size

A1 $\quad f$ is bounded below, that is, $f(x) \geq f_{\mathrm{low}}$ for all x.

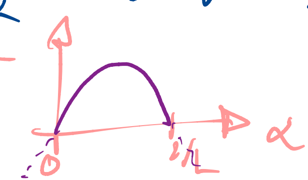
A2 $\quad f\big(R_x(s)\big) \leq f(x) + \langle \mathrm{grad} f(x), s\rangle_x + \frac{L}{2}\|s\|_x^2$ for all $(x,s) \in \mathrm{T}\mathcal{M}$.

Theorem: With $\alpha \in (0, 2/L)$, gradient descent finds small gradients.

Proof. $\quad x_{k+1} = R_{x_k}\big(\overbrace{-\alpha\, \mathrm{grad} f(x_k)}^{s}\big)$

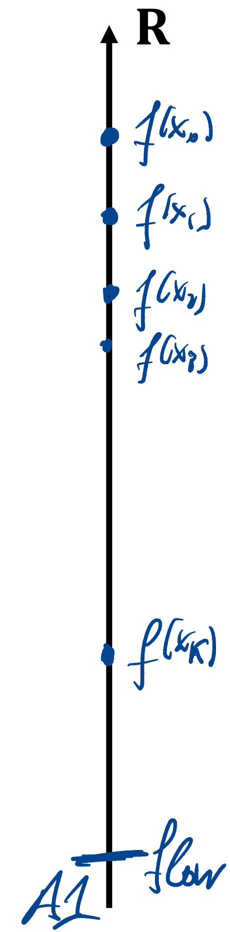
$$\underline{A2} \Rightarrow f(x_{k+1}) \leq f(x_k) + \langle \mathrm{grad} f(x_k), -\alpha\, \mathrm{grad} f(x_k)\rangle_{x_k} + \frac{L}{2}\alpha^2 \|\mathrm{grad} f(x_k)\|_{x_k}^2$$

$$f(x_k) - f(x_{k+1}) \geq \left(\alpha - \frac{\alpha^2 L}{2}\right) \|\mathrm{grad} f(x_k)\|_{x_k}^2$$



$< \qquad \underbrace{\qquad\qquad}_{>0 \ \forall \alpha \in (0,\, 2/L).}$

$$
\begin{aligned}
f(x_0) - f_{low} &\overset{A1}{\geq} f(x_0) - f(x_K) \\
&= \sum_{k=0}^{K-1} f(x_k) - f(x_{k+1}) \\
&\overset{A2}{\geq} \sum_{k=0}^{K-1} c \, \|\operatorname{grad} f(x_k)\|_{x_k}^2 \\
&\geq K c \min_{0 \leq k \leq K-1} \|\operatorname{grad} f(x_k)\|_{x_k}^2
\end{aligned}
$$



Beyond constant step sizes

A1 f is bounded below, that is, $f(x) \geq f_{\text{low}}$ for all x.

A3 Sufficient decrease: $f(x_k) - f(x_{k+1}) \geq c\|\operatorname{grad} f(x_k)\|_{x_k}^2 \quad \forall k$.

Theorem: Under **A1**, any sequence verifying **A3** with $c > 0$ enjoys:

if any

$$\lim_{k\to\infty} \|\operatorname{grad} f(x_k)\|_{x_k} = 0$$ i.e., accumulation points are critical.

$$\|\operatorname{grad} f(x_k)\|_{x_k} \leq \sqrt{\frac{f(x_0) - f_{\text{low}}}{c}} \frac{1}{\sqrt{K}}$$ for all K and some $k < K$.

This is traditionally referred to as a "global convergence".

A word about the regularity assumption

A2 $\quad f\big(R_x(s)\big) \le f(x) + \langle \mathrm{grad} f(x), s\rangle_x + \frac{L}{2}\|s\|_x^2$ for all $(x, s) \in \mathrm{T}\mathcal{M}$.

① If $\mathcal{M} = \mathcal{E}$, and $R_x(s) = x + s$, then

$$A2 \equiv \quad f(x+s) \le f(x) + \langle \mathrm{grad} f(x), s\rangle + \frac{L}{2}\|s\|^2 \quad \forall x, s \in \mathcal{E}$$

$$A2 \Leftarrow \mathrm{grad} f \text{ is } L\text{-Lipschitz continuous}$$

$$\Leftarrow \|\mathrm{Hess} f(x)\| \le L \text{ for all } x.$$

② If $\mathcal{M}$ is Riemannian and $R = \mathrm{Exp}$, then all of ① extends.

We call **A2** a Lipschitz-type assumption. More in textbook §4.4, §10.4.