Worst-case complexity bounds for optimization on manifolds

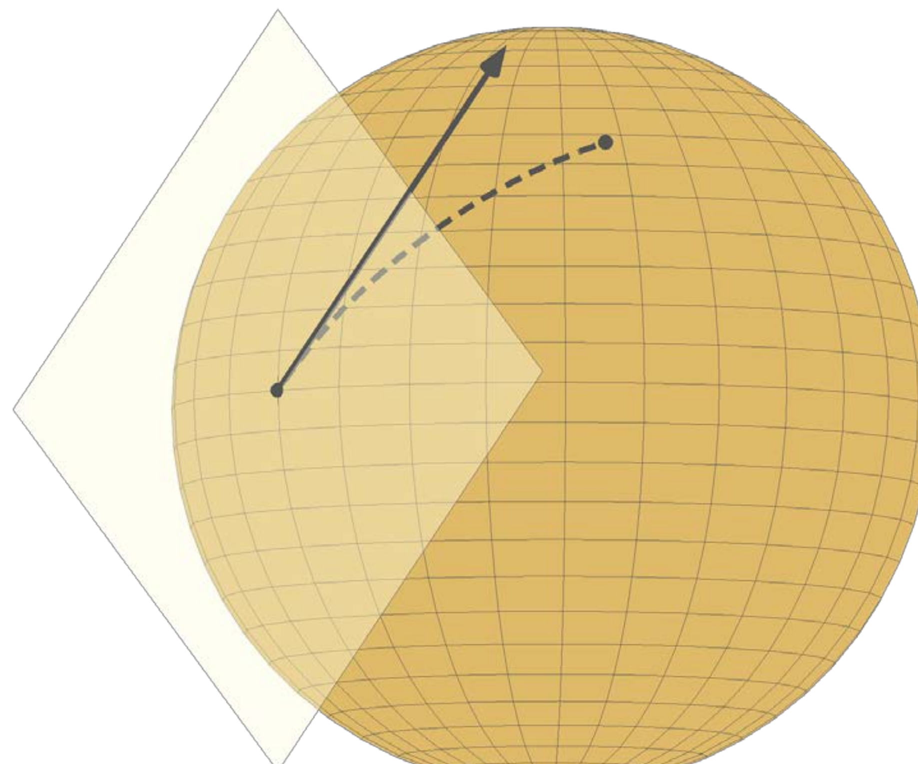
Nicolas Boumal
Princeton University

with Naman Agarwal, Brian Bullins, Coralia Cartis and Pierre-Antoine Absil

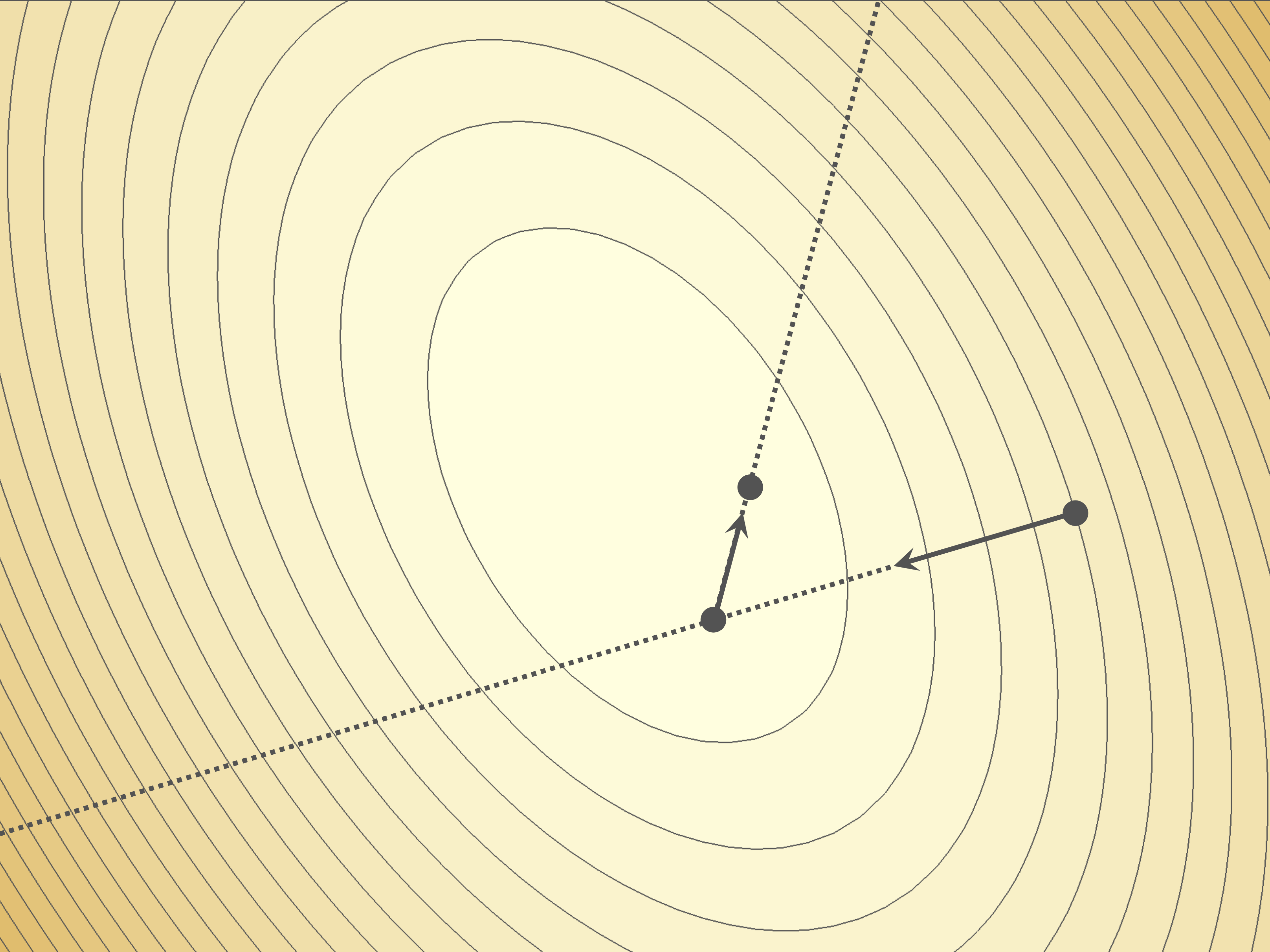


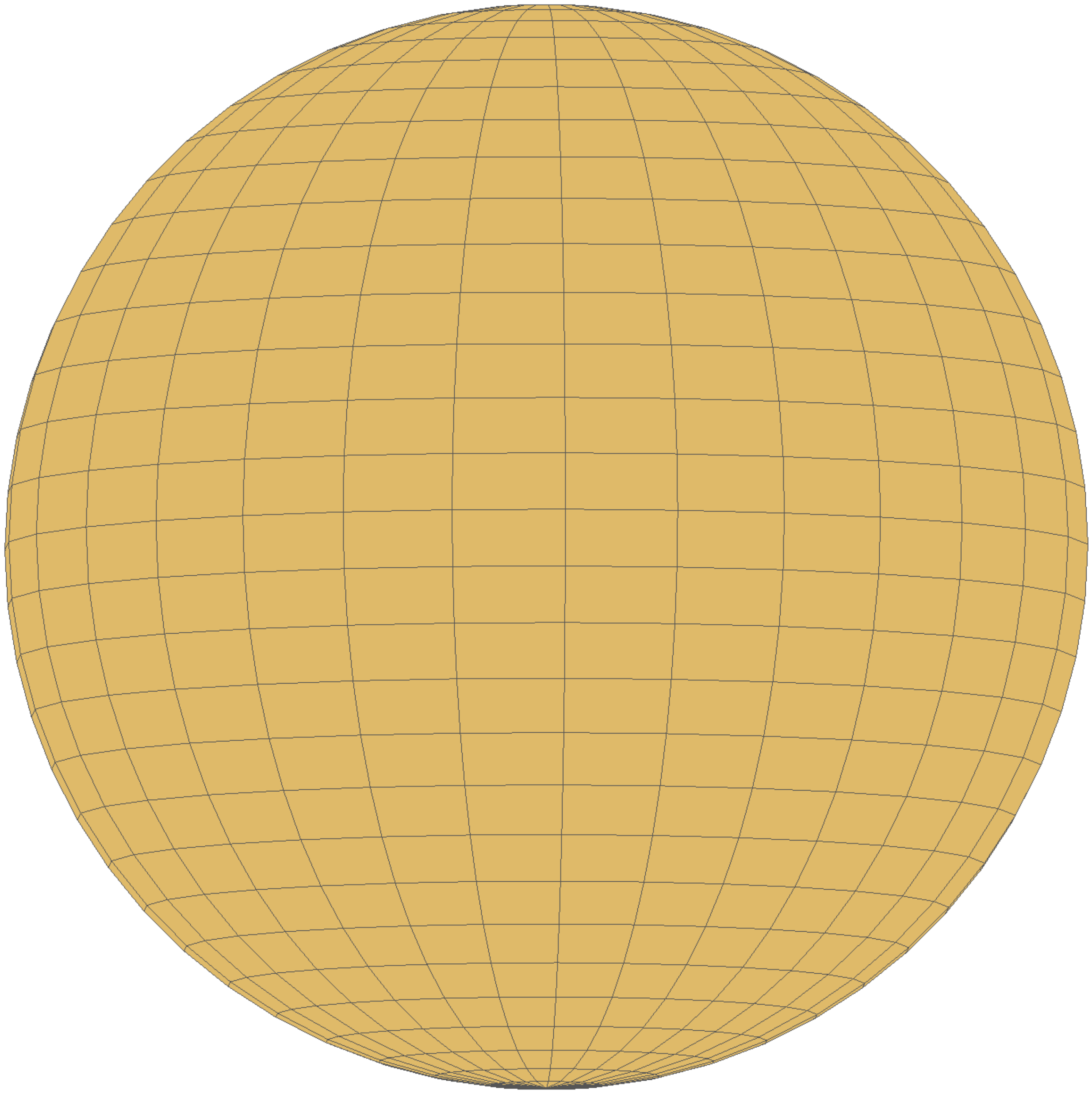

Optimization on manifolds

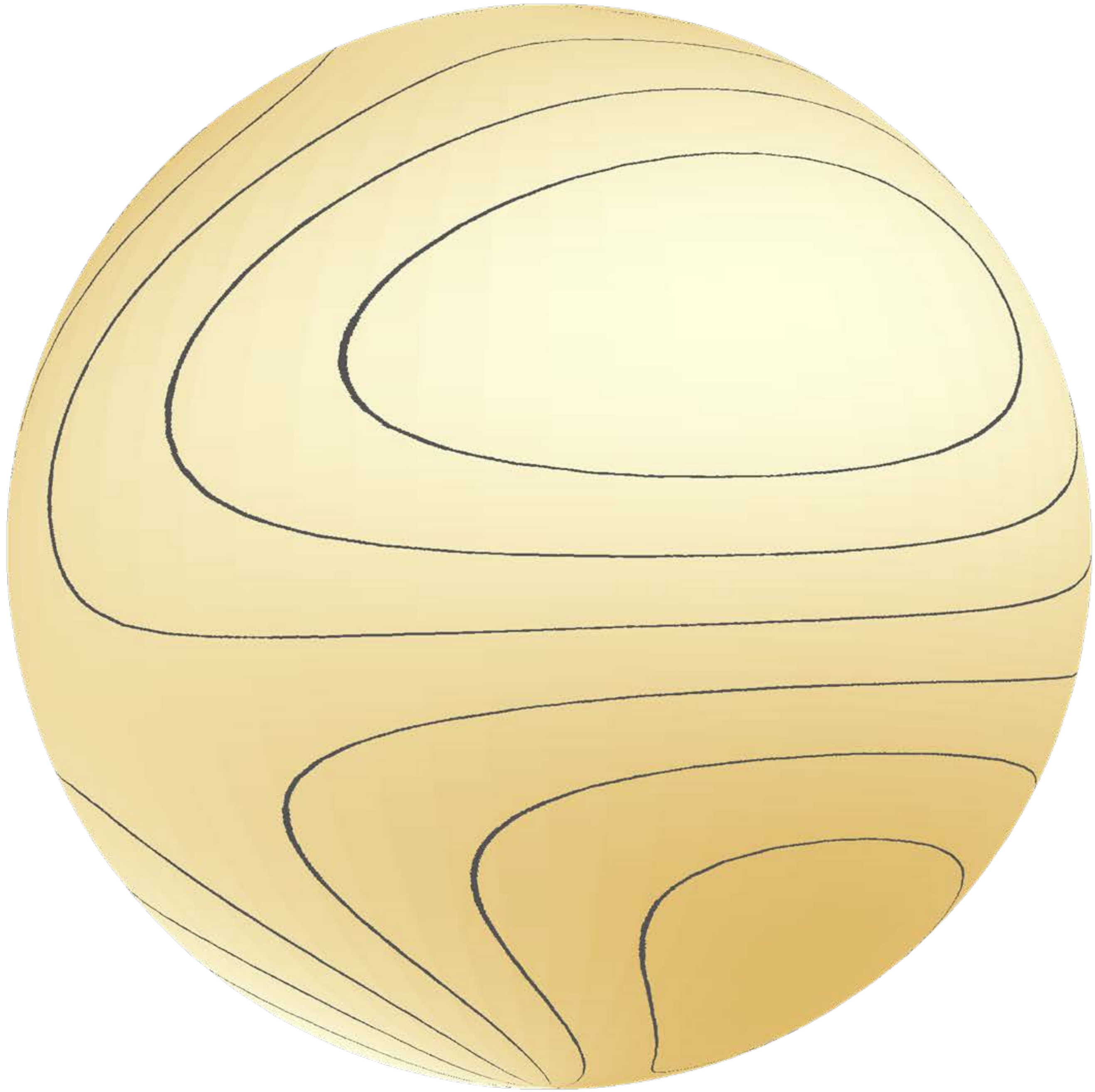
$$\min_{x \in \mathcal{M}} f(x)$$

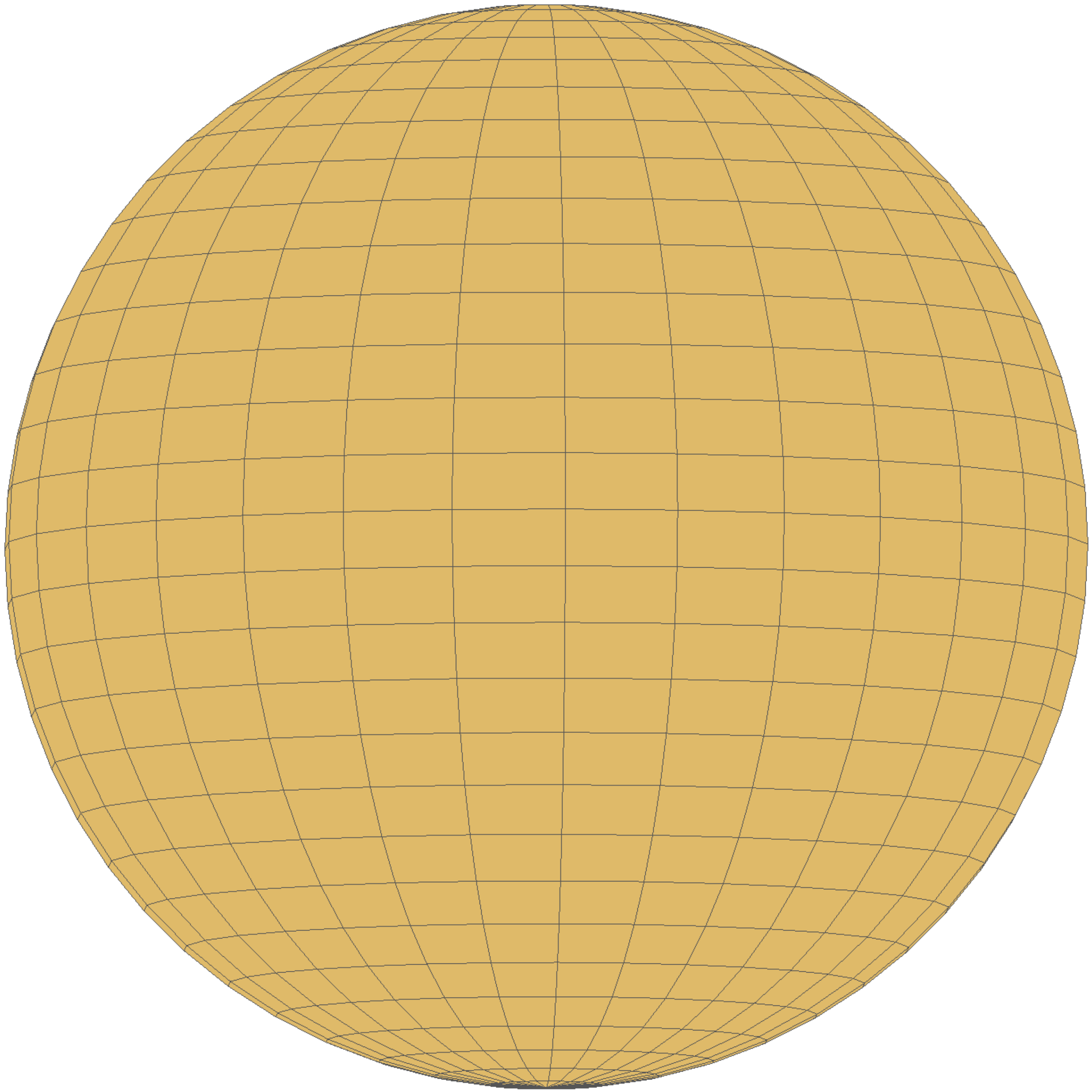


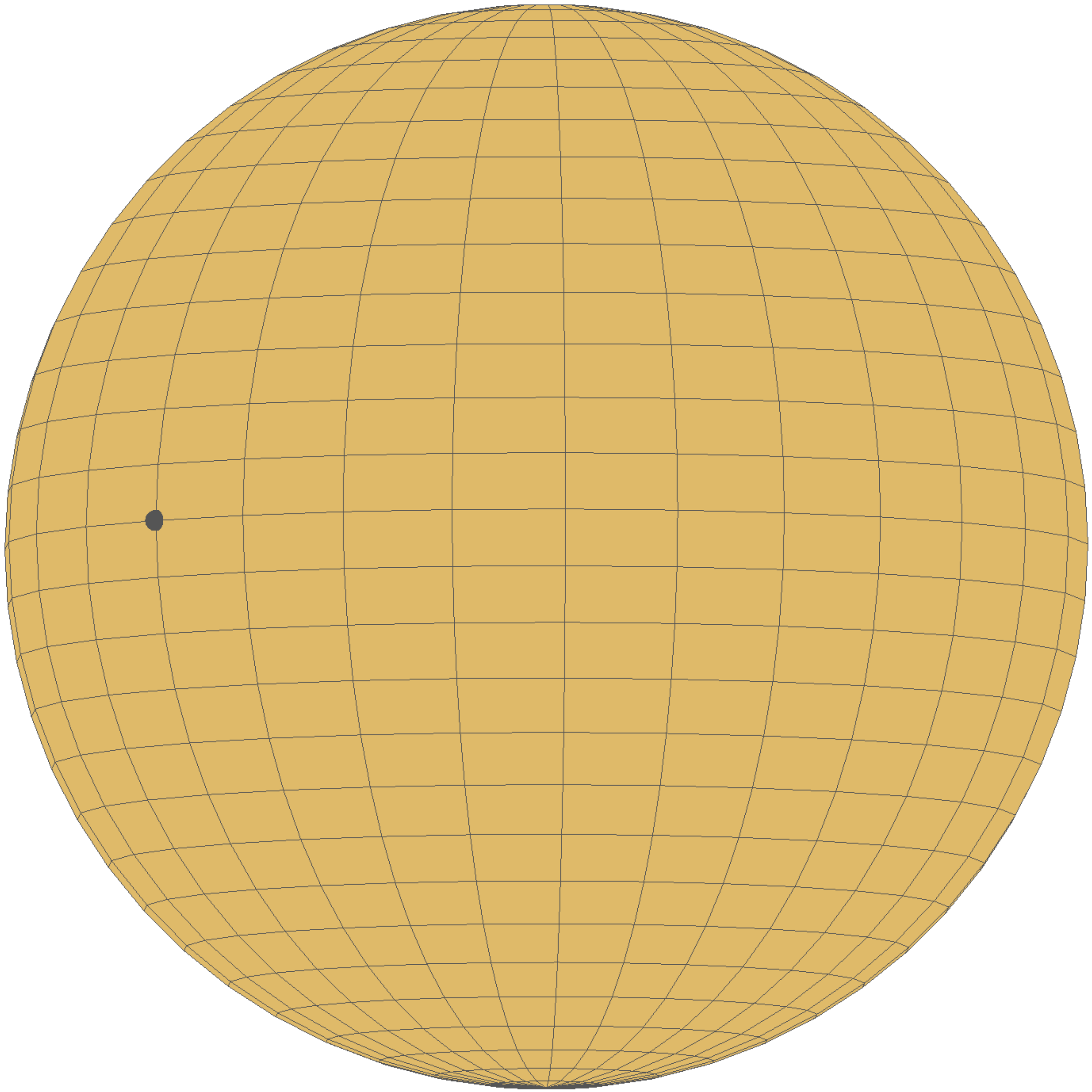
Taking a close look at gradient descent

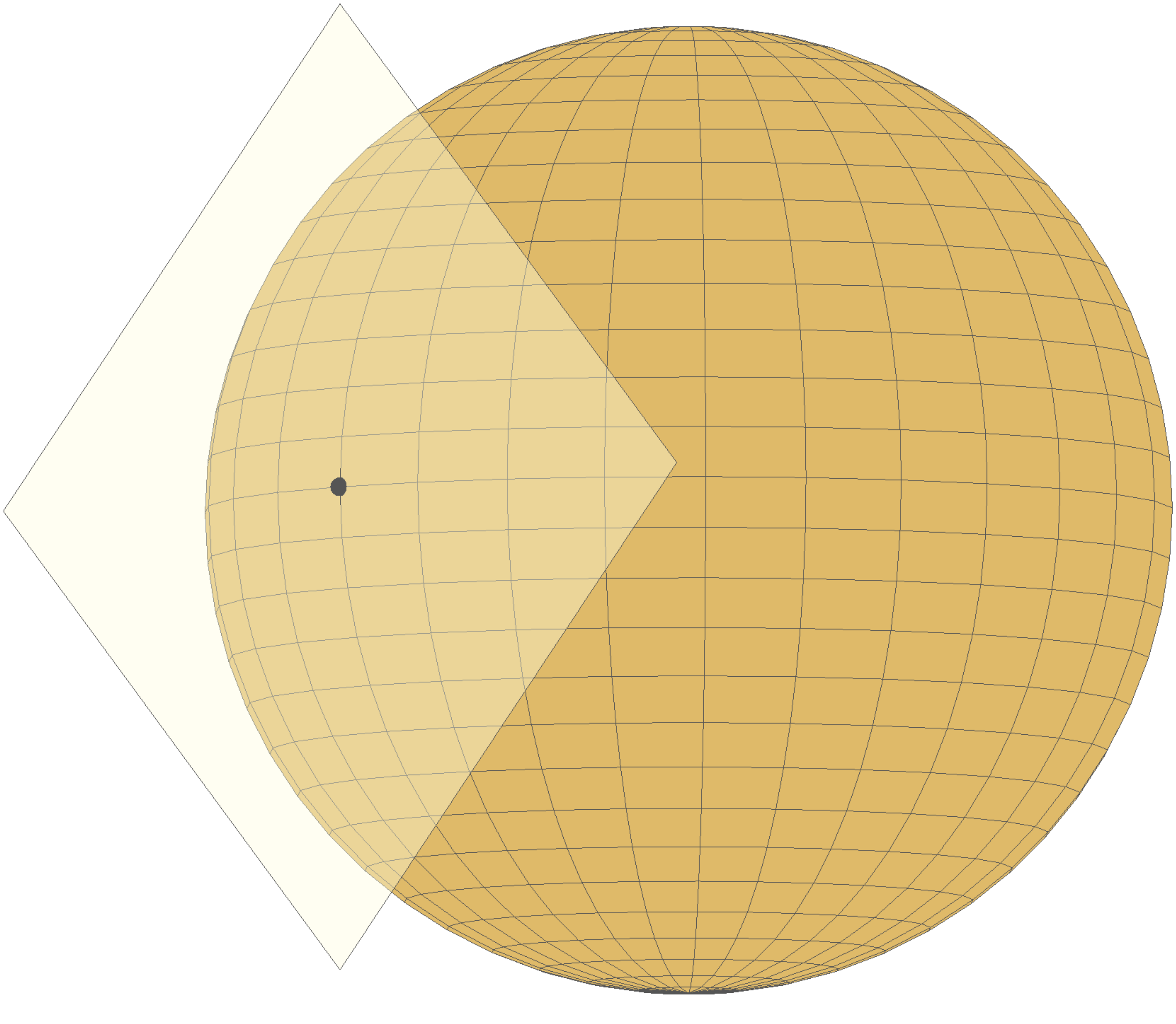


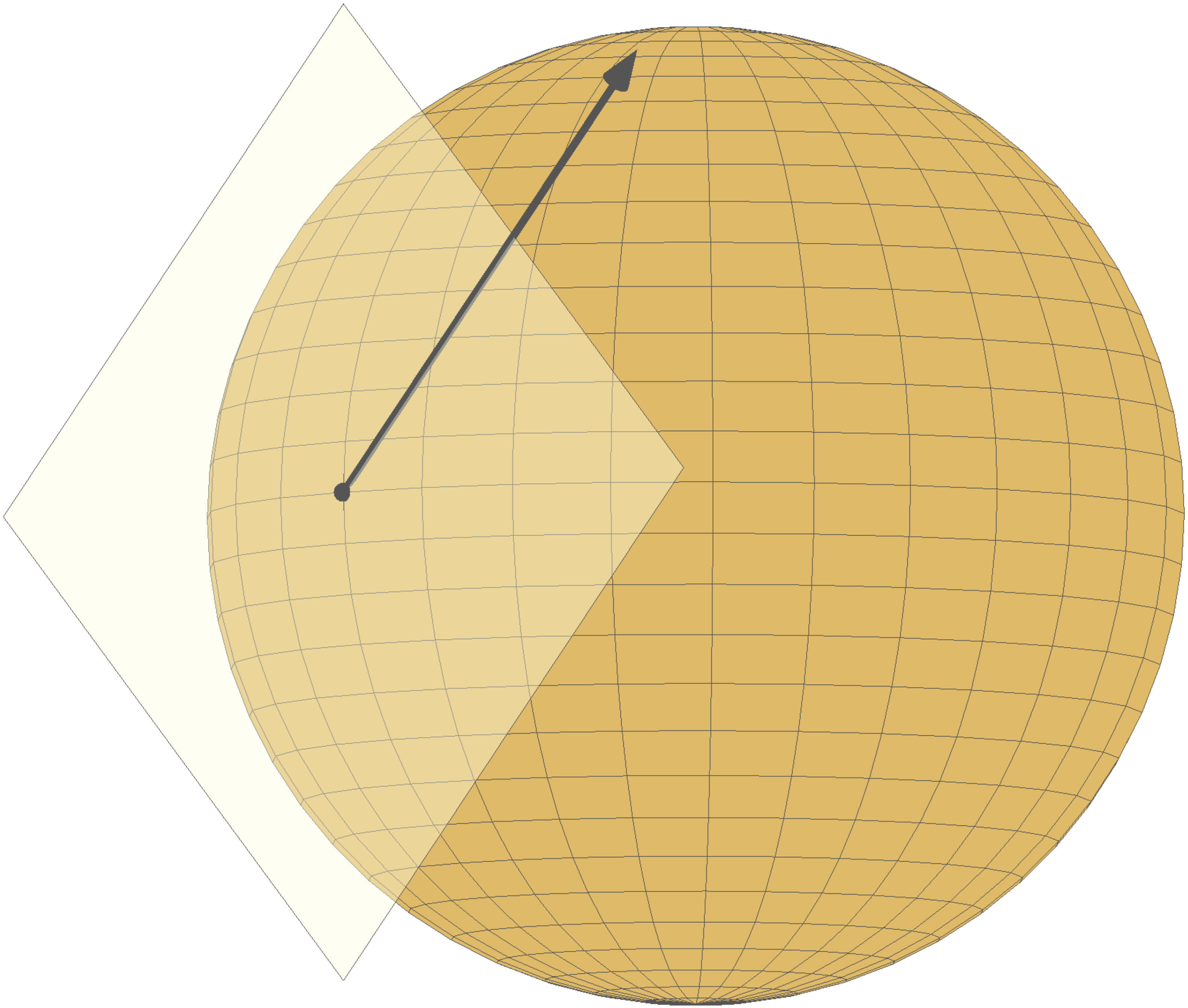


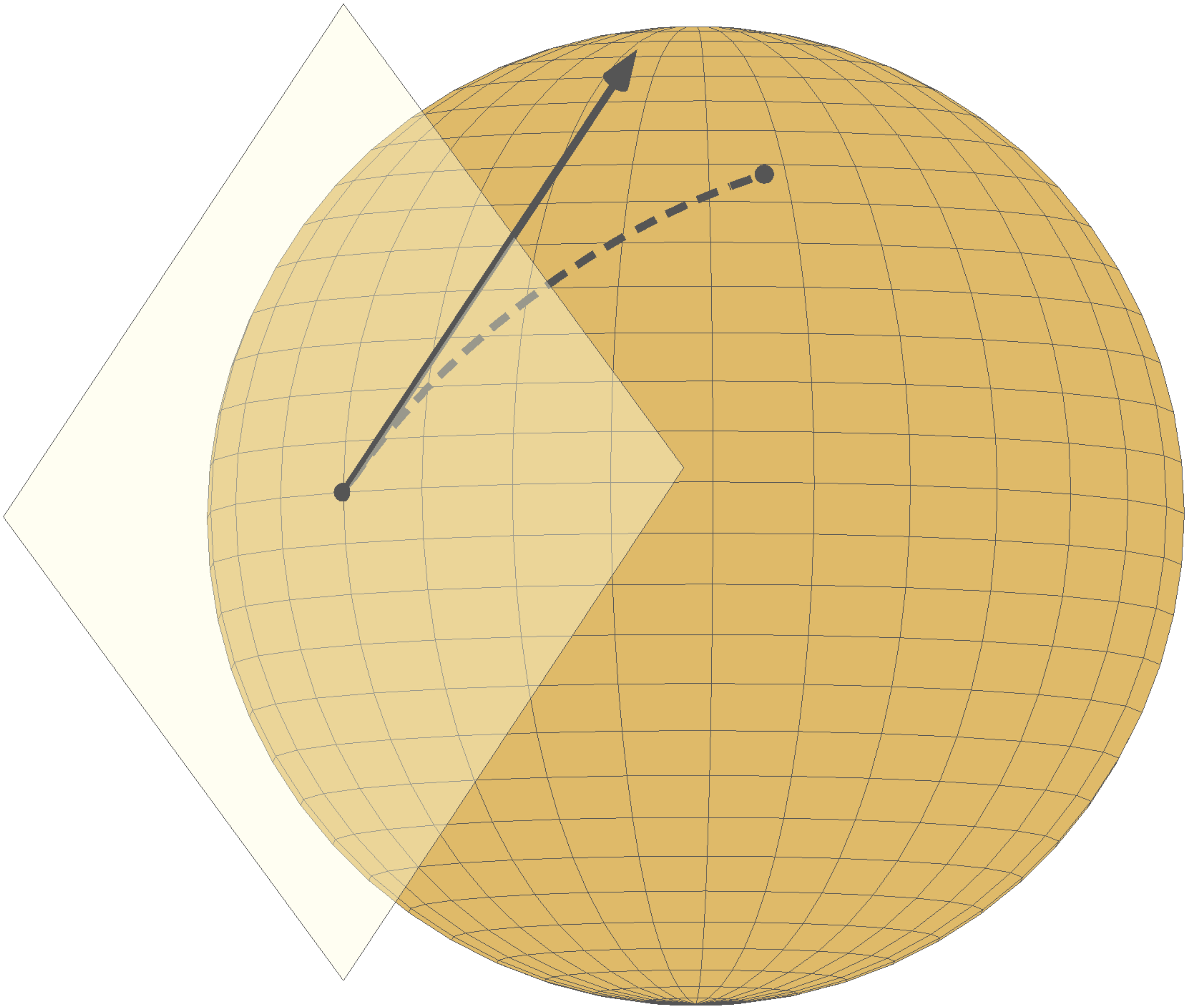












Gifts from the smooth & Riemannian structure

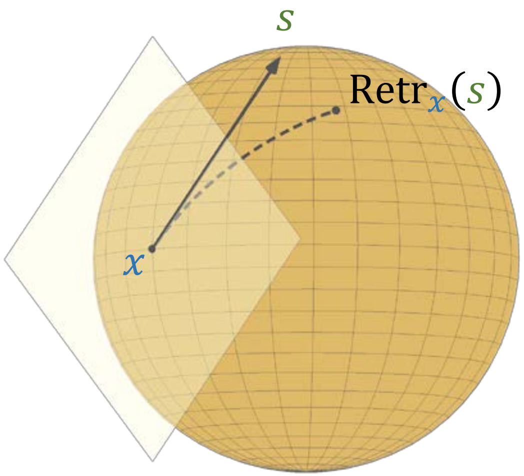
$$\min_{x \in \mathcal{M}} f(x)$$

Tangent spaces: allowed directions

E.g.: $\mathrm{T}_x\mathcal{M} = \{s \in \mathbf{R}^n : x^\mathrm{T} s = 0\}$

Retractions: tools to move around

E.g.: $\mathrm{Retr}_x(s) = \frac{x+s}{\|x+s\|}$



Inner products: gradient, Hessian

E.g.: $\langle s_1, s_2 \rangle_x = s_1^\mathrm{T} s_2$

These ideas have been around since the 70s (Luenberger, Gabay)

Gradient descent in $\mathbf{R}^n$

A1 $f(x) \geq f_{\text{low}}$ for all $x \in \mathbf{R}^n$

A2 ∇f is L-Lipschitz: $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$

Algorithm: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

Complexity: $\|\nabla f(x_K)\| \leq \varepsilon$ for some $K \leq 2L(f(x_0) - f_{\text{low}})\frac{1}{\varepsilon^2}$

How do we generalize **A2** to manifolds?

- A proper Lipschitz definition is inconvenient:

$$\mathrm{dist}\big(\mathrm{grad} f(y), \mathrm{grad} f(x)\big) \leq L \cdot \mathrm{dist}(x, y)$$

- Opportunistic approach: extract what we need from the proof.

Gradient descent in $\mathbf{R}^n$

A1 $f(x) \geq f_{\text{low}}$ for all $x \in \mathbf{R}^n$

A2 ∇f is L-Lipschitz: $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$

Algorithm: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

Complexity: $\|\nabla f(x_K)\| \leq \varepsilon$ for some $K \leq 2L(f(x_0) - f_{\text{low}})\frac{1}{\varepsilon^2}$

$$\mathbf{A2} \Rightarrow |f(y) - f(x) - \langle y - x, \nabla f(x)\rangle| \leq \frac{L}{2}\|y - x\|^2$$

$$\Rightarrow f(x_{k+1}) - f(x_k) + \frac{1}{L}\langle \nabla f(x_k), \nabla f(x_k)\rangle \leq \frac{1}{2L}\|\nabla f(x_k)\|^2$$

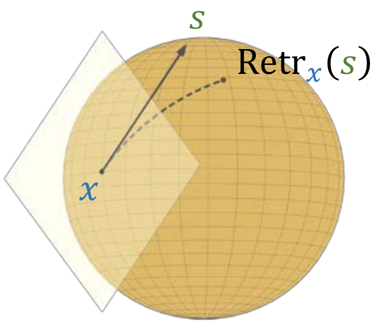
$$\Rightarrow f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$\mathbf{A1} \Rightarrow f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{K} f(x_k) - f(x_{k+1}) > \frac{\varepsilon^2}{2L}(K+1)$$

Gradient descent on $\mathcal{M}$

A1 $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$

A2 $f(\text{Retr}_x(s)) - f(x) - \langle s, \text{grad}f(x) \rangle \leq \frac{L}{2}\|s\|^2$



Algorithm: $x_{k+1} = \text{Retr}_{x_k}\left(-\frac{1}{L}\text{grad}f(x_k)\right)$

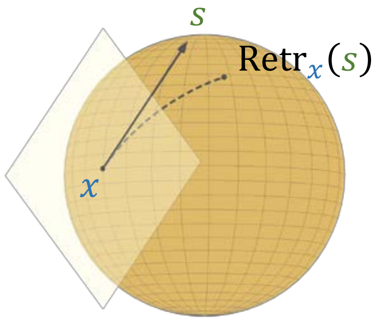
Complexity: $\|\text{grad}f(x_K)\| \leq \varepsilon$ with $K \leq 2L(f(x_0) - f_{\text{low}})\frac{1}{\varepsilon^2}$

$$\textbf{A2} \Rightarrow f(x_{k+1}) - f(x_k) + \frac{1}{L}\|\text{grad}f(x_k)\|^2 \leq \frac{1}{2L}\|\text{grad}f(x_k)\|^2$$

$$\Rightarrow f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\text{grad}f(x_k)\|^2$$

$$\textbf{A1} \Rightarrow f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{K} f(x_k) - f(x_{k+1}) > \frac{\varepsilon^2}{2L}(K+1)$$

$$\textbf{A2} \quad f\big(\mathrm{Retr}_x(s)\big) - f(x) - \langle s, \mathrm{grad} f(x)\rangle \leq \frac{L}{2}\|s\|^2$$



Assumption on both f and Retr.

Satisfied *in particular*:

1. If $\mathcal{M} \subset \mathbf{R}^n$ is compact and ∇f is locally Lipschitz in $\mathbf{R}^n$.

2. If $\mathcal{M}$ is compact and Retr is "nice".

Ongoing research.

Beyond gradient descent on manifolds

Trust regions

arXiv:1605.08101

$O\left(\varepsilon^{-2}\right)$ for small gradient
$O\left(\varepsilon^{-3}\right)$ for second-order too


Adaptive regularization with cubics (ARC)

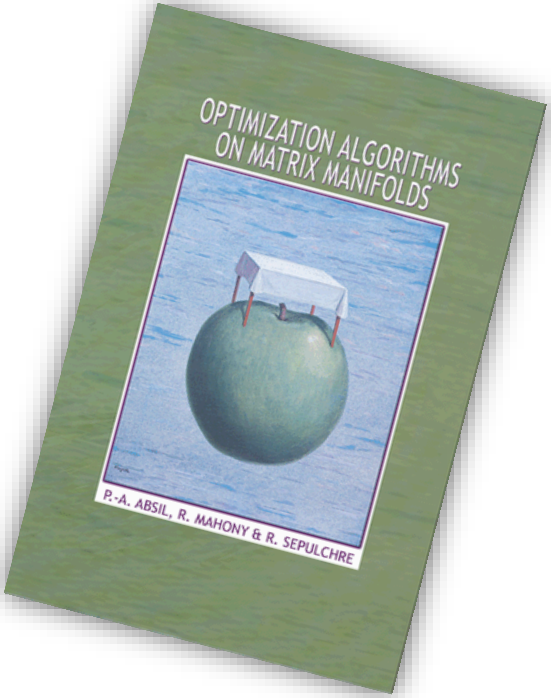
arXiv:1806.00065


$O\left(\varepsilon^{-1.5}\right)$ for small gradient
$O\left(\varepsilon^{-3}\right)$ for second-order too

See also Zhang & Zhang's work on cubics: arXiv:1805.05565

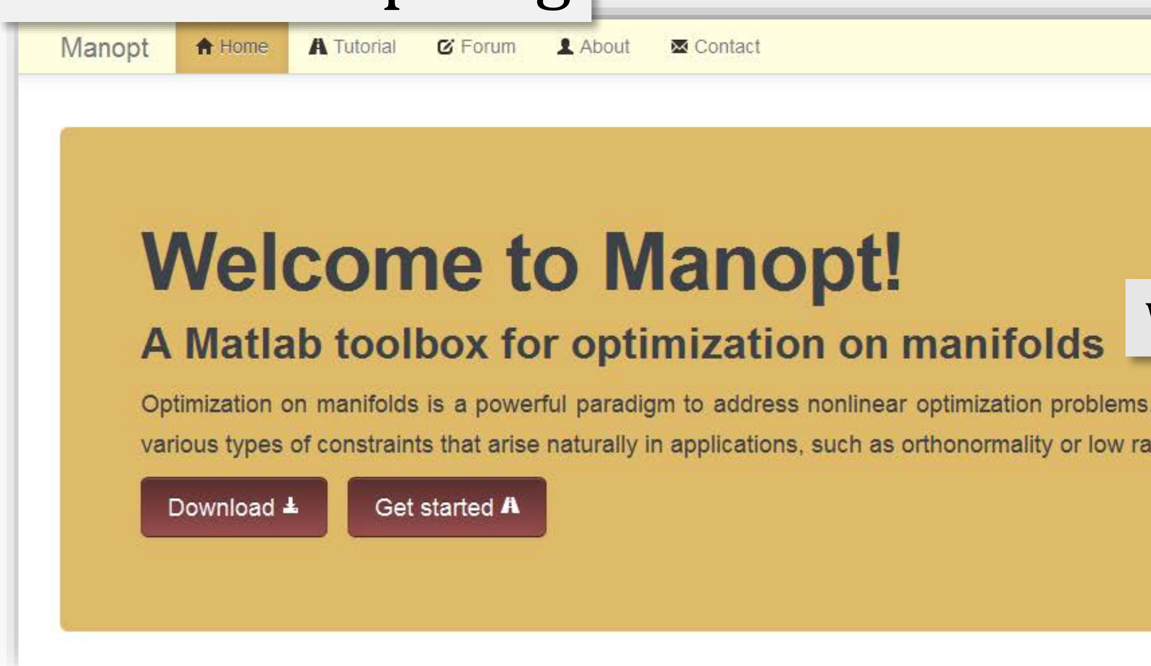
An excellent book

Optimization algorithms on matrix manifolds



A Matlab toolbox

www.manopt.org



With Mishra, Absil & Sepulchre