Optimization flows landing on the Stiefel manifold

Simon Vary*

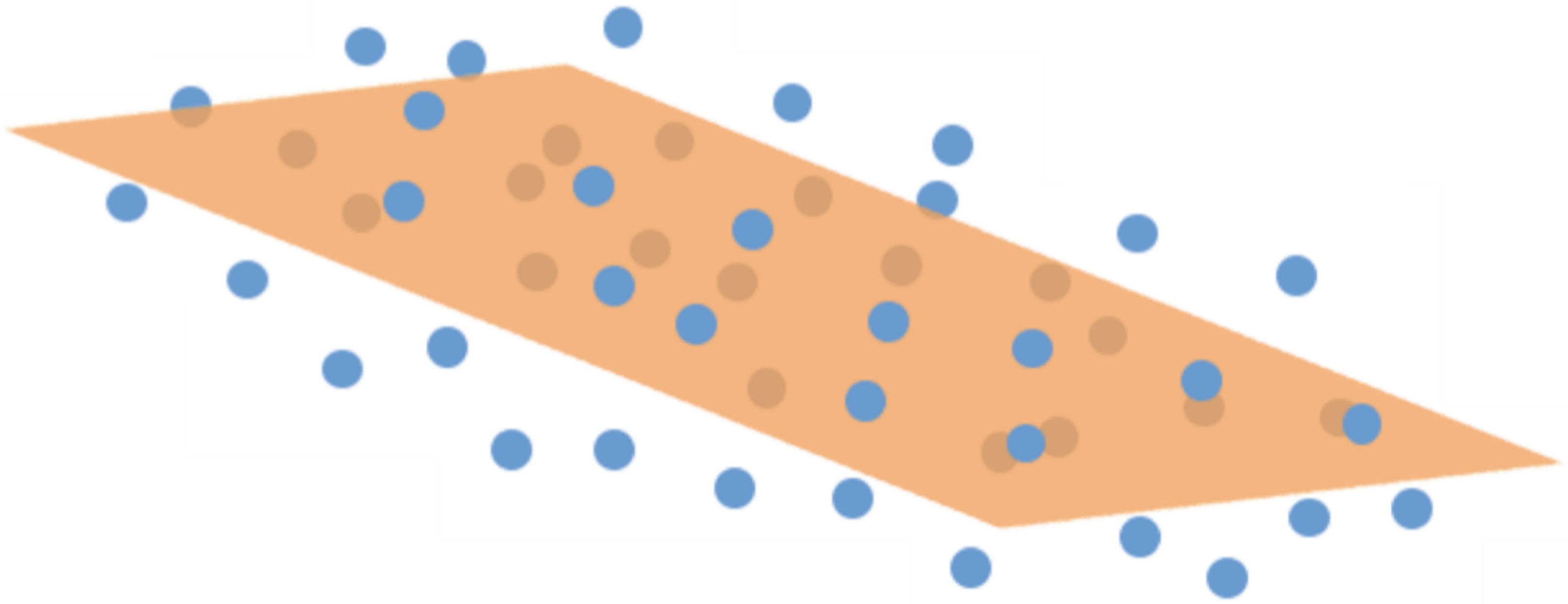
joint work with Bin Gao, Pierre Ablin***, Pierre-Antoine Absil***

** ICTEAM, Université catholique de Louvain*
*** Institute for Applied Mathematics, University of Münster*
**** CNRS, Université Paris-Dauphine, PSL University*
Machine Learning Group, Apple, Paris

25th International Symposium on Mathematical Theory of Networks and Systems MTNS 2022 12/09/22

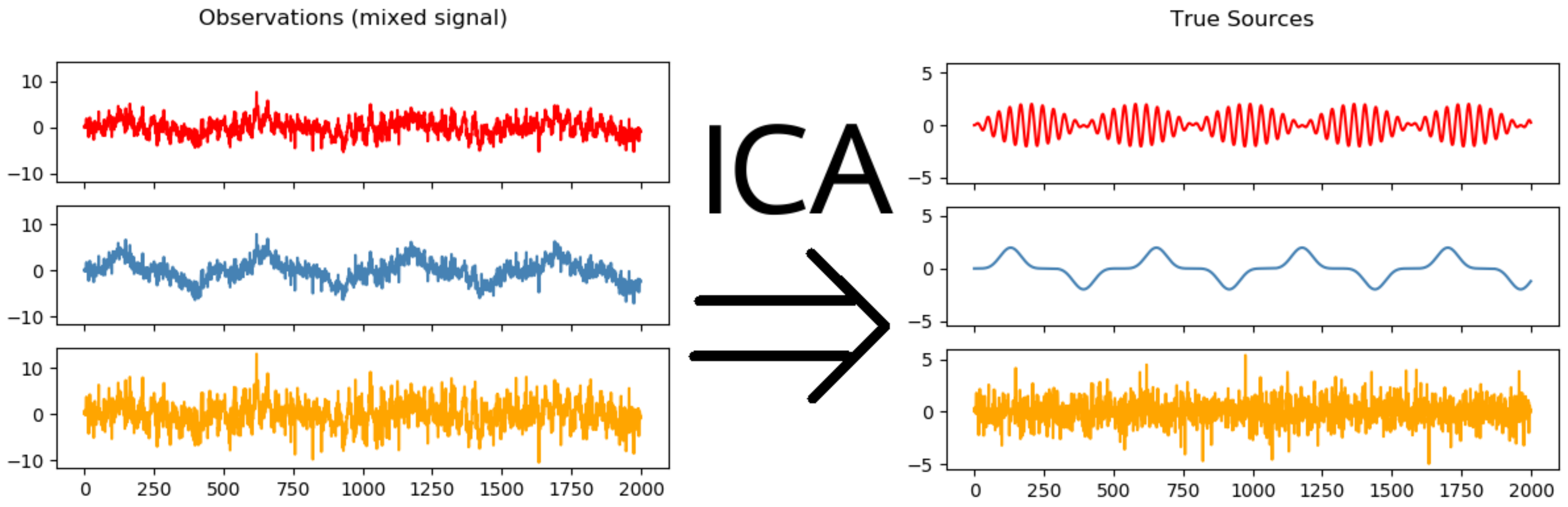
Optimization over the Stiefel manifold

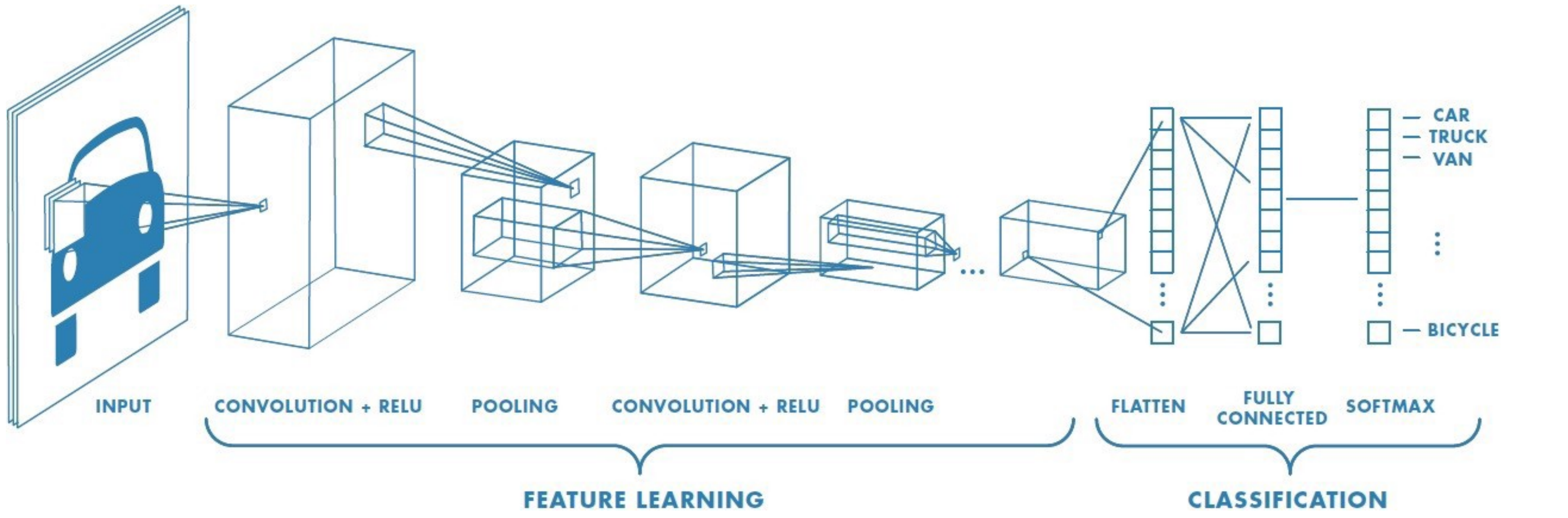
$$\min_{X \in \mathbb{R}^{n\times p}} f(x) \quad \text{s.t.} \quad \mathrm{St}(p,n) := \left\{ X \in \mathbb{R}^{n\times p} : X^\top X = I_p \right\}$$



Applications such as

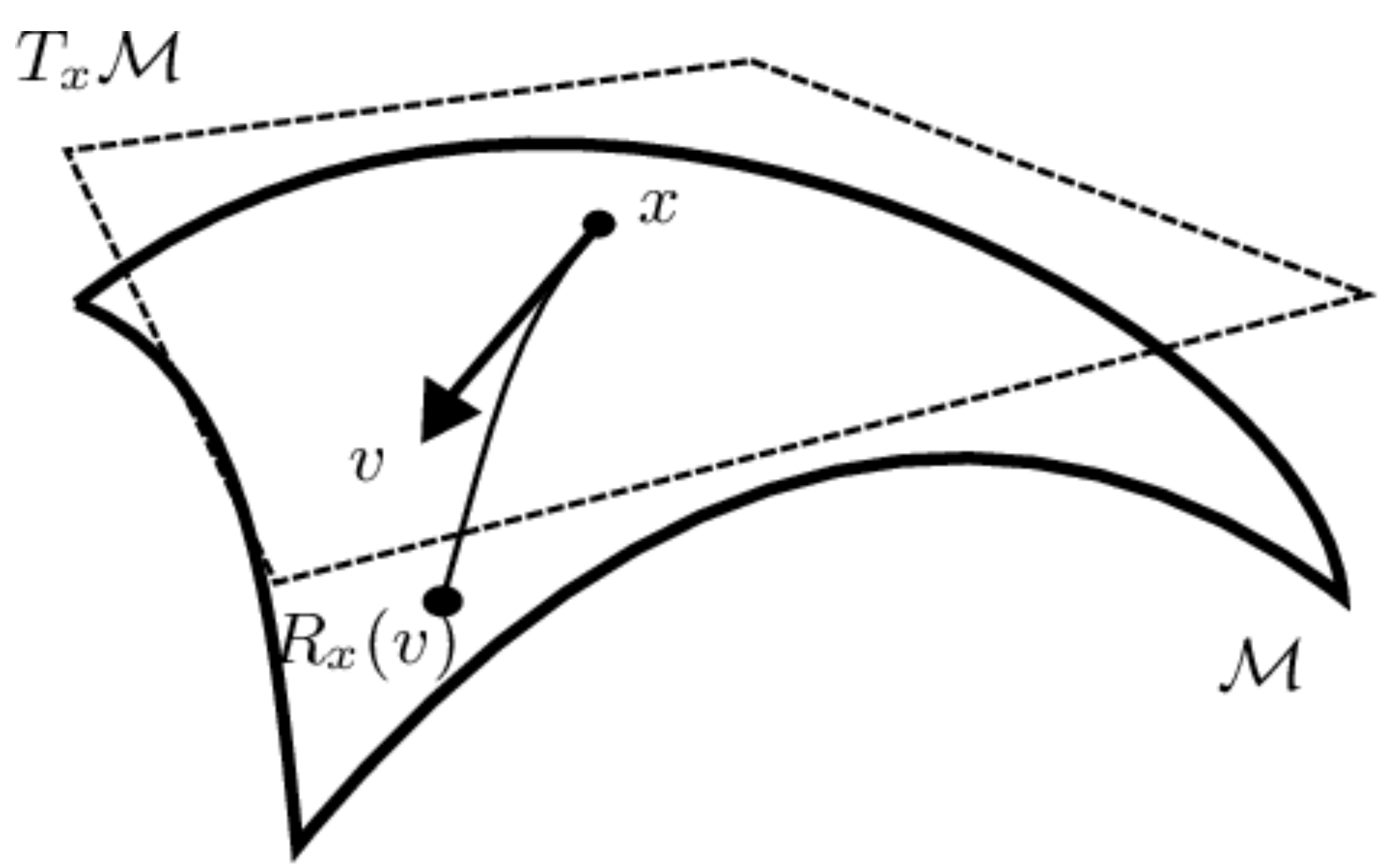
- Principal component analysis $\qquad \max_{X \in \mathrm{St}(p,n)} \frac{1}{2} \| AX \|_F^2$
- Independent component analysis
- Orthogonal weights in deep learning





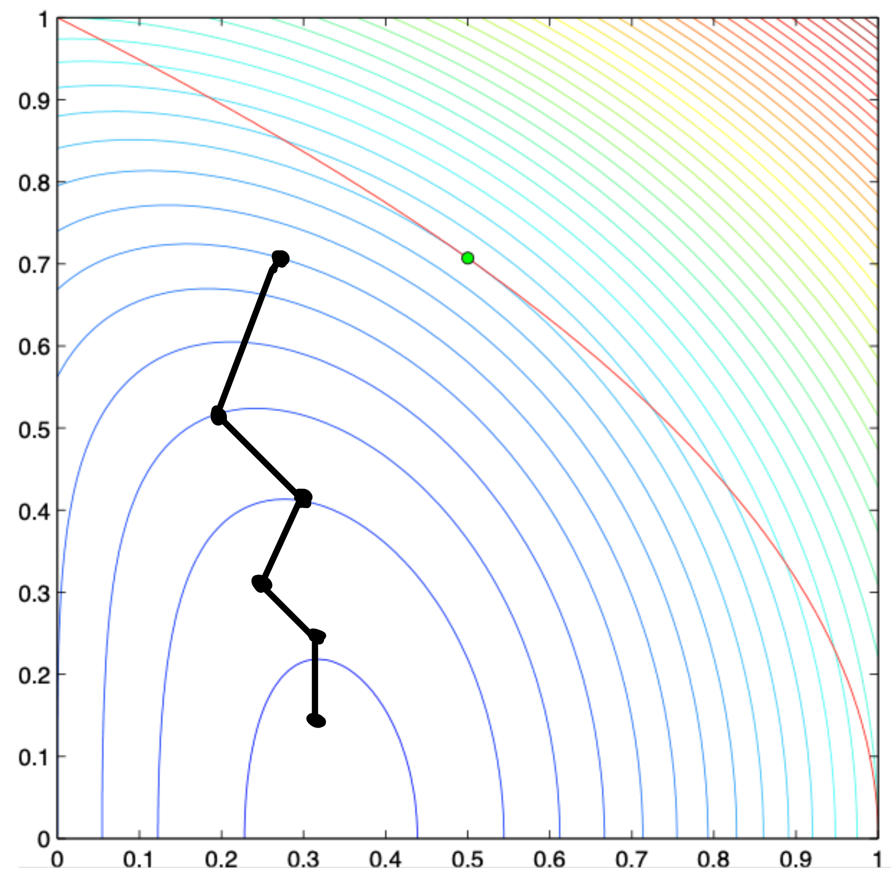
Optimization approaches

Riemmanian optimization

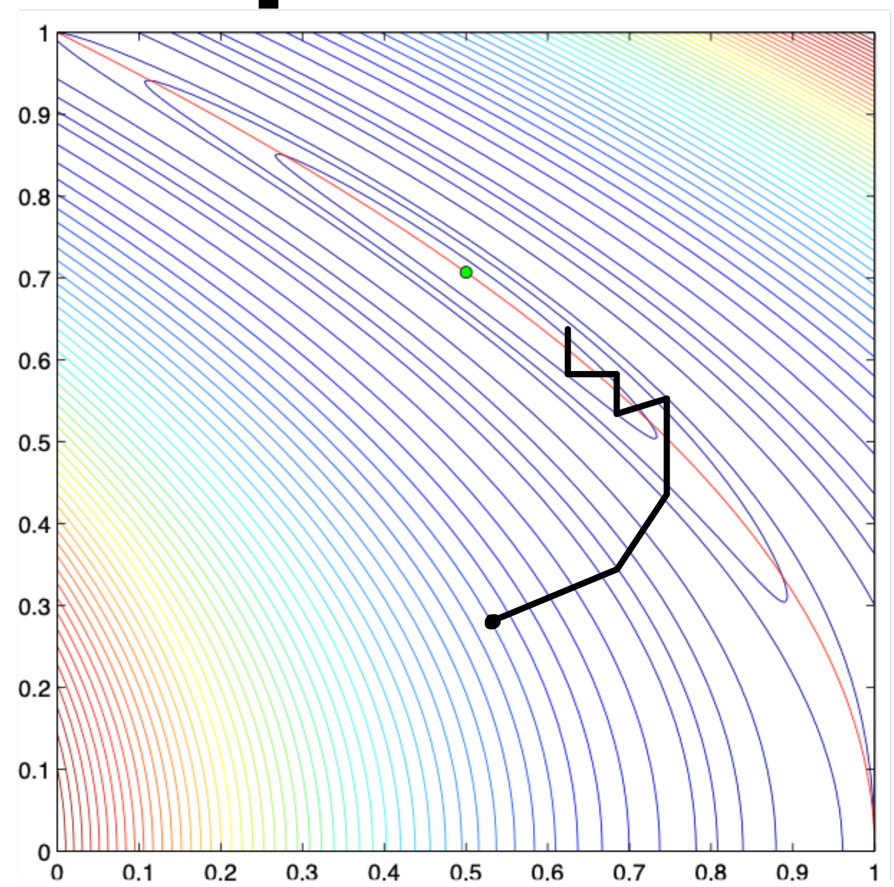


- $X^{t+1} = R_{X^t}(-\operatorname{grad} f(X^t))$
- Retraction: QR $\mathcal{O}(np^2)$, Cayley $\mathcal{O}(p^3)$
- $\operatorname{grad} f(X) = \operatorname{skew}(\nabla f(X)X^\top)X \quad (\mathcal{O}(np^2))$

Infeasible constrained optimization



$\sigma = 1$



$\sigma = 0.01$

from lecture slides of Coralia Cartis

- Penalty methods: $+\frac{1}{4\sigma}\|X^\top X - I_p\|_F^2$
- Augmented Lagrangian Method
- Adaptively choosing parameters can be tricky
- Gradient of the penalty: $X(X^\top X - I_p)$ (mat. mult. $\mathcal{O}(np^2)$)

Landing field

- Consider the following flow

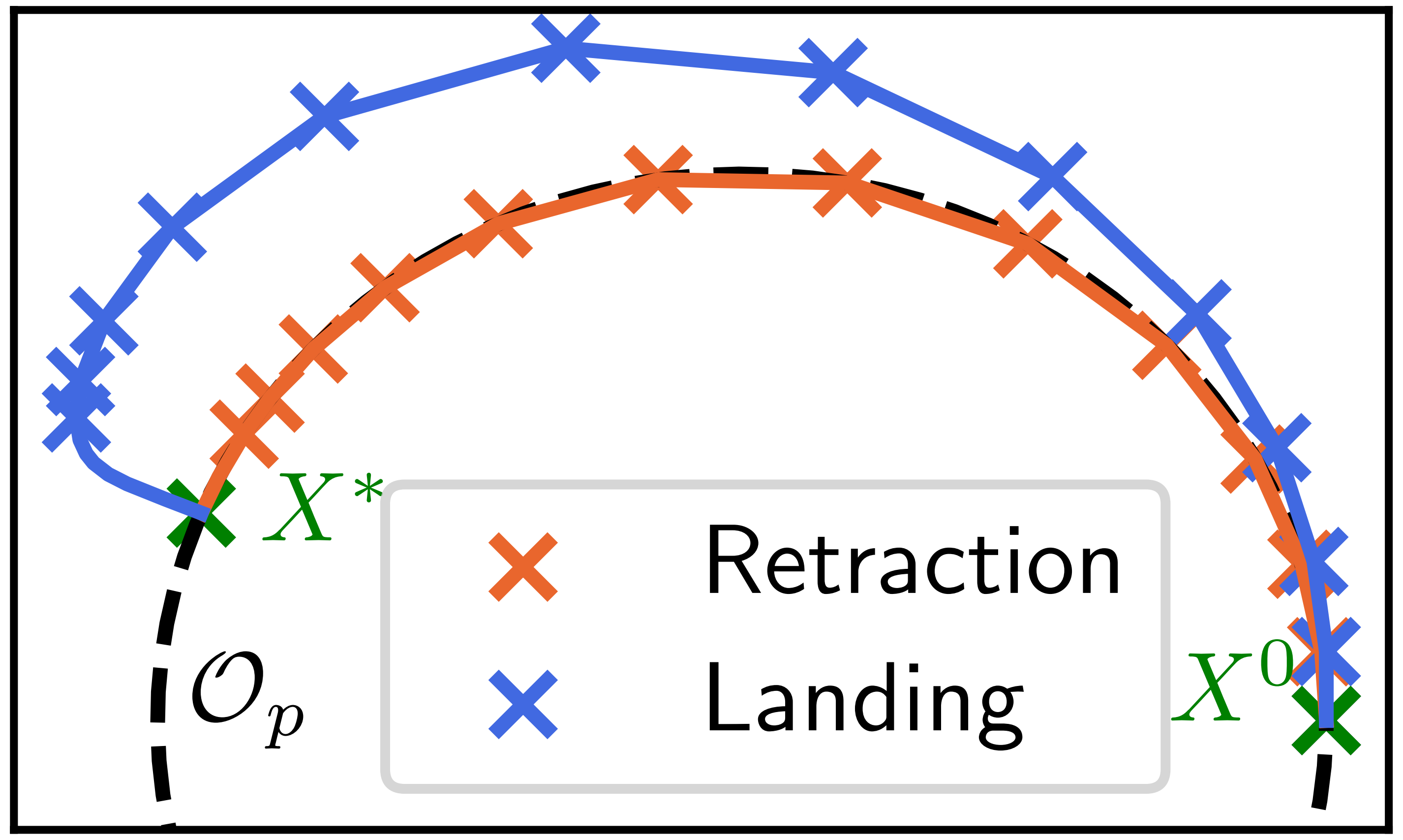
$$\dot{X}(t) = -\Lambda\left(X(t)\right),$$

where

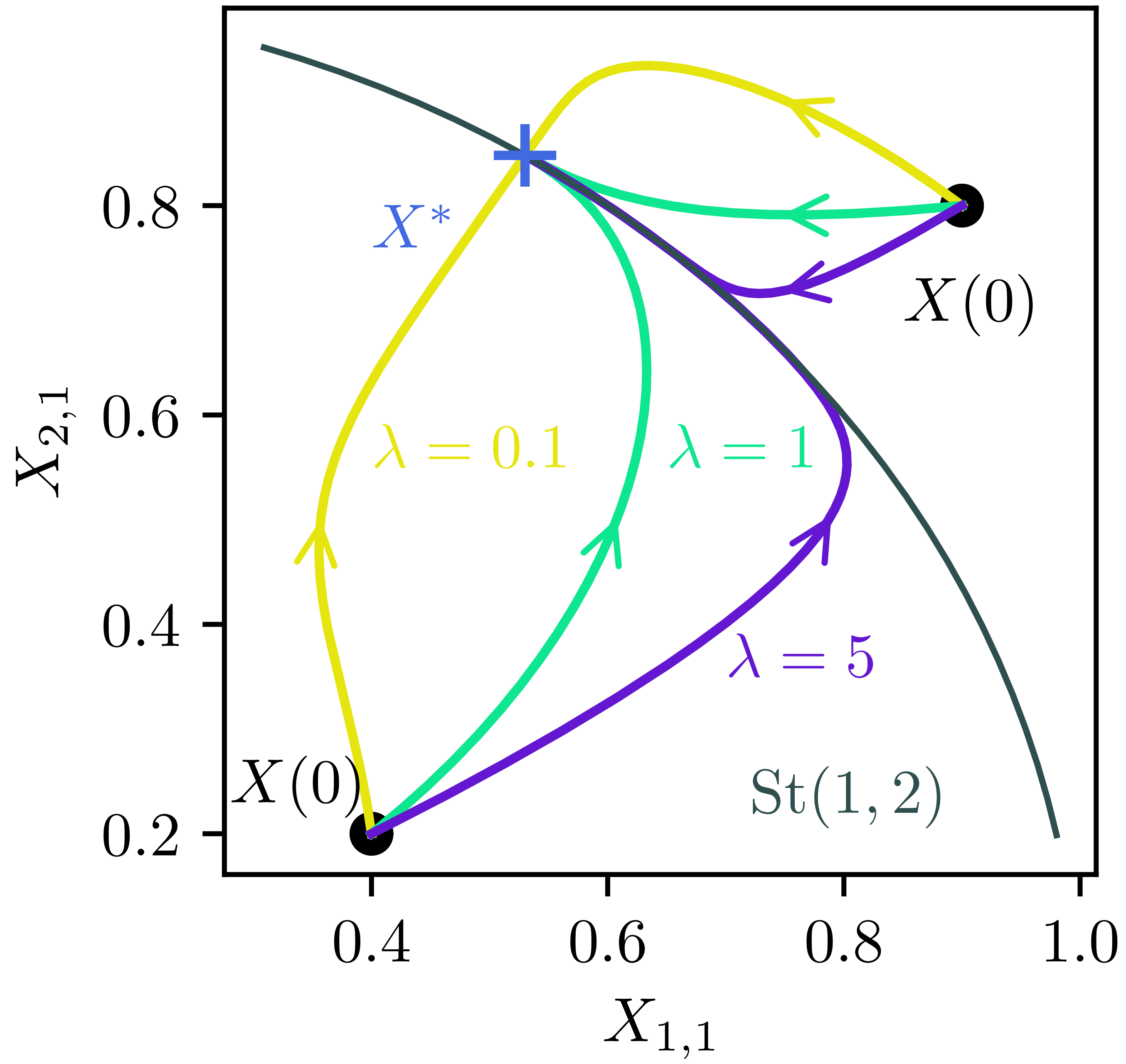
$$\Lambda(X) := \psi(X)X + \lambda \nabla \mathcal{N}(X),$$

and

- $\nabla \mathcal{N}(X) = X(X^\top X - I_p)$, with $\mathcal{N}(X) := \frac{1}{4}\left\| X^\top X - I_p \right\|_F^2$,
- $\psi(X) = 2\mathrm{skew}\left(\nabla f(X) X^\top\right)$, with $\mathrm{skew}(A) = \frac{1}{2}\left(A - A^\top\right)$



Fast and accurate optimization on the orthogonal manifold.
Ablin & Peyré, AISTATS 2022



Interpretation of the landing field

(on the Stiefel manifold)

$$\Lambda(X) := \psi(X)X + \lambda \nabla \mathcal{N}(X)$$
$$= 2\mathrm{skew}\left(\nabla f(X)X^\top\right)X + \lambda X(X^\top X - I_p)$$

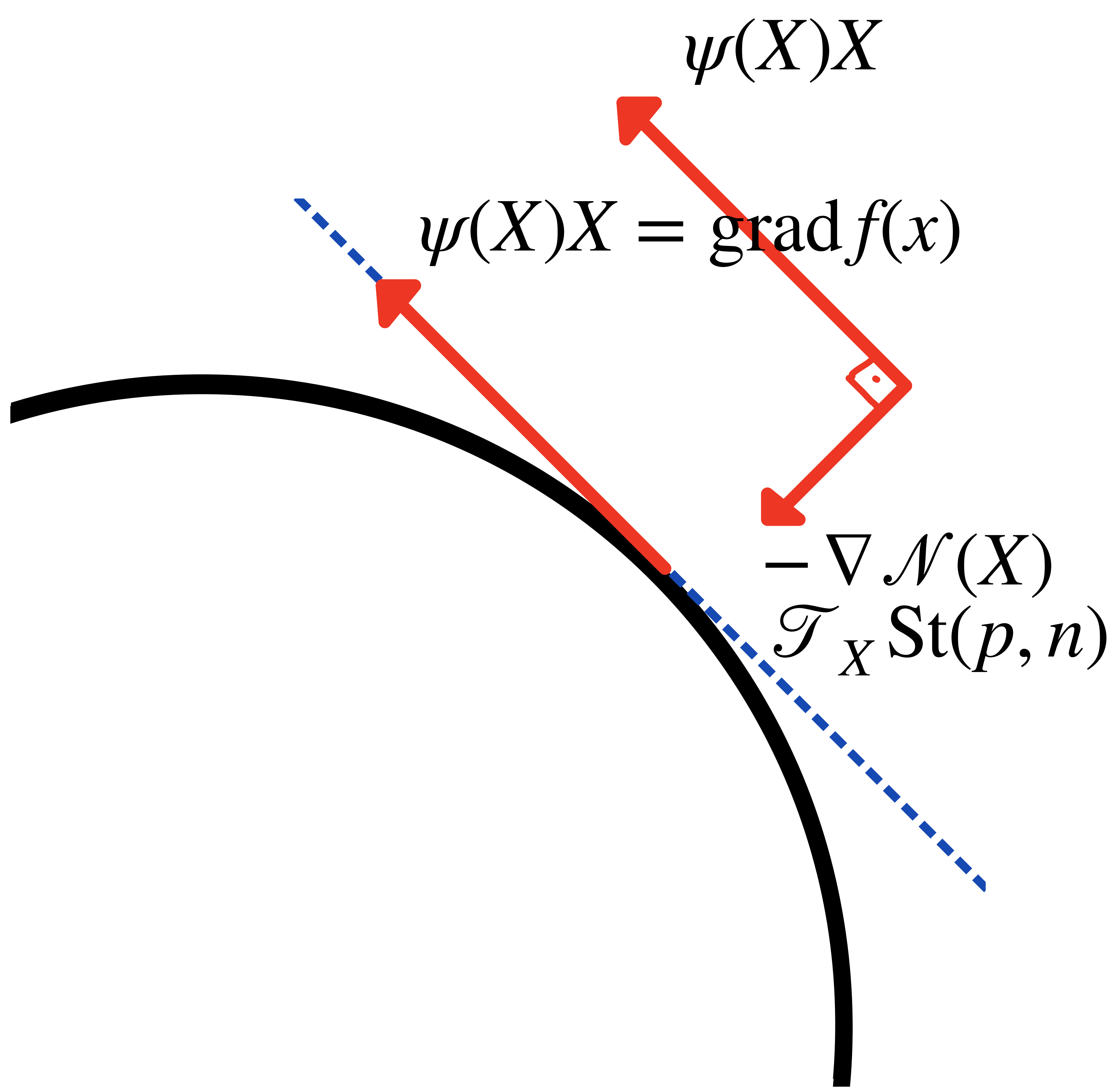
- Orthogonality

$$\langle \psi(X)X, X(X^\top X - I_p\rangle = 0$$

- Interpretation of $\psi(X)X$ in the canonical metric $g_X^{\mathrm{c}}(\cdot,\cdot)$*

$$\psi(X)X = \mathrm{Proj}_{\mathcal{T}_X \mathrm{St}(p,n)} \nabla f(X) = \mathrm{grad} f(X)$$
$$g_X^{\mathrm{c}}(\xi,\zeta) := \langle \xi, (I_n - \frac{1}{2}XX^\top)\zeta\rangle \quad \text{for all } \xi,\zeta \in \mathbb{R}^{n\times p},$$

- What is the interpretation when away from the Stiefel manifold?



* The Geometry of Algorithms with Orthogonality Constraints. Edelman et. al, SIAM. J. Matrix Anal. & Appl. 1998

UCLouvain

Interpretation of the landing field

(away from the Stiefel manifold)

- Consider a generalization of the Stiefel manifold and a map

$$\forall M \succ 0 : \mathrm{St}_M(p,n) := \{Y \in \mathbb{R}^{n\times p} : Y^\top Y = M\} \quad \text{and} \quad \Phi_M : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p} : X \mapsto Y = XM^{\frac{1}{2}}.$$

- We can define a metric

$$g_Y(\xi,\zeta) := g^{\mathrm{c}}_{\Phi_M^{-1}(Y)}\left(\Phi_M^{-1}(\xi), \Phi_M^{-1}(\zeta)\right) \quad \text{where} \quad g^{\mathrm{c}}_X(\xi,\zeta) := \langle \xi, (I_n - \frac{1}{2}XX^\top)\zeta\rangle$$

Φ_M isometry

$$(\mathrm{St}(p,n), g^{\mathrm{c}}) \longleftrightarrow (\mathrm{St}_M(p,n), g)$$

Interpretation of the landing field

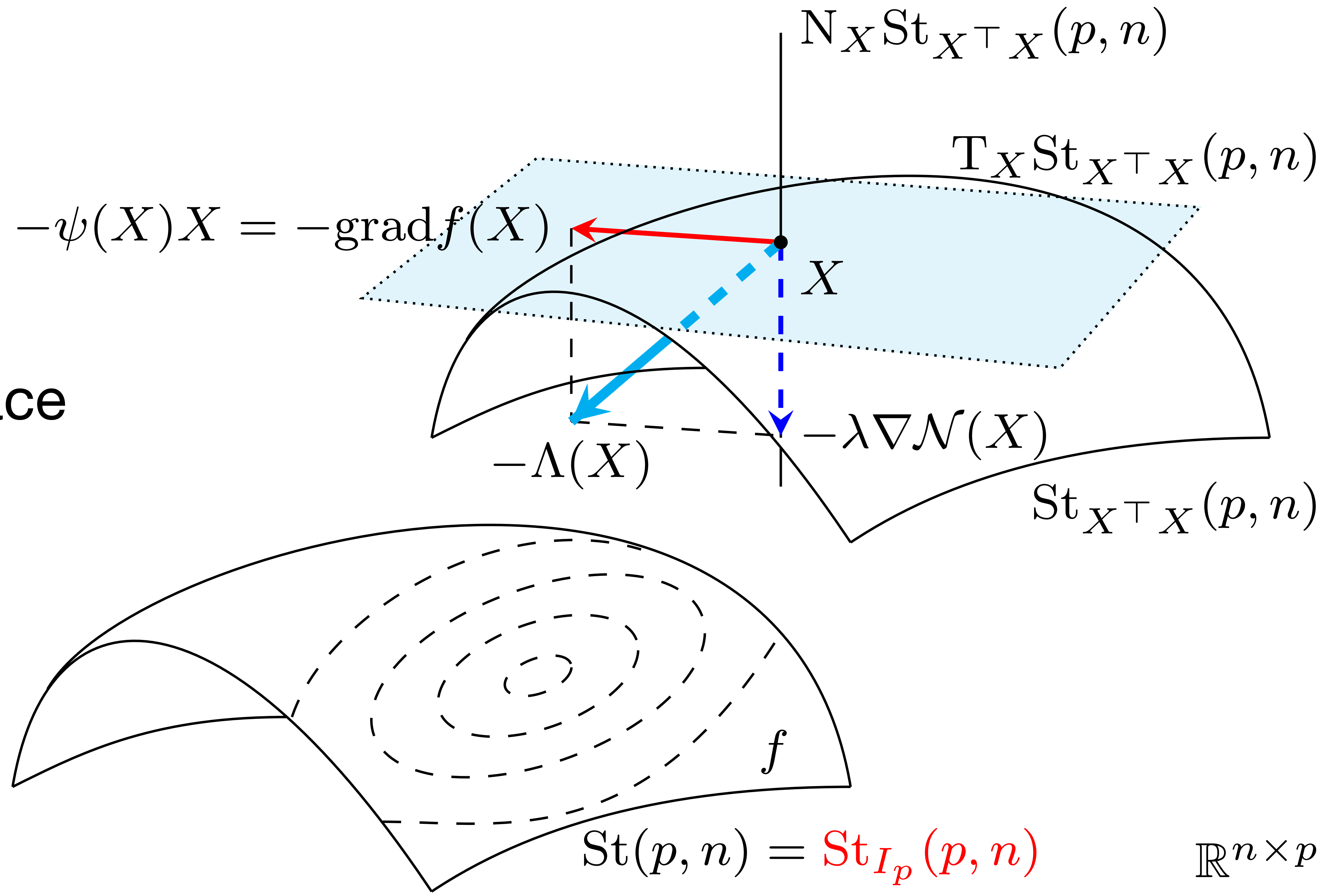
(on the generalized Stiefel manifold $\mathrm{St}_{X^\top X}(p,n)$)

- Riemannian gradient of f on $\left(\mathrm{St}_{X^\top X}(p,n), g_{X^\top X}\right)$

$$\mathrm{grad}_{\mathrm{St}_{X^\top X}(p,n)} f(X) = \psi(X)X$$

- The normal component belongs to the normal space

$$\nabla \mathcal{N}(X) \in \mathrm{N}_X\mathrm{St}_{X^\top X}(p,n)$$



Convergence of the landing flow

(existence and uniqueness)

Differentiate $\mathcal{N}\left(X(t)\right)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{N}(X(t)) = \langle \dot{X}(t), \nabla \mathcal{N}(X(t))\rangle$$

by orthogonality

$$= -\langle \psi(X(t))X(t) + \lambda \nabla \mathcal{N}(X(t)), \nabla \mathcal{N}(X(t))\rangle$$

$$= -\lambda \|\nabla \mathcal{N}(X(t))\|_F^2 \leq 0, \qquad \text{for any} \quad \lambda > 0$$

- hence $\mathcal{N}\left(X(t)\right)$ is <u>nonincreasing</u> and remains in a <u>closed compact</u> set,
- for $\nabla f(X)$ locally Lipschitz, $\Lambda(X)$ is also locally Lipschitz and,
- by Picard–Lindelöf theorem, there is a unique solution $t \mapsto \varphi_t(X_0)$ such that $\varphi_0(X_0) = X_0$

Convergence of the landing flow

(convergence to the Stiefel manifold)

- Let $\chi(t) = X(t)^\top X(t)$

$$\dot{\chi}(t) = \dot{X}(t)^\top X(t) + X(t)^\top \dot{X}(t), \qquad \text{where} \quad \dot{X}(t) = -\Lambda\left(X(t)\right) = \psi\left(X(t)\right) X(t) + \lambda \nabla \mathcal{N}\left(X(t)\right)$$

$$= -2\lambda\, \chi(t)\Big(\chi(t) - I_p\Big),$$

- polynomial of a symmetric matrix => $\chi(t)$ has constant eigenvectors

- The eigenvalues of $\chi(t)$ evolve as:

$$\chi_i(t) = \frac{\chi_i(0)e^{2\lambda t}}{\chi_i(0)(e^{2\lambda t} - 1) + 1} \quad \longrightarrow \quad 1, \qquad \text{for} \quad \lambda > 0, \quad \text{thus} \quad \lim_{t\to\infty} \mathcal{N}(\phi_t(X_0)) = 0.$$

Convergence of the landing flow
(convergence to the local minima)

- Let $\mathscr{C} \subseteq \mathrm{St}(p, n)$ be the set of critical points on the manifold, we have

  $X^* \in \mathscr{C}$ if and only if $\Lambda(X^*) = 0,$ by $\psi(X)X$ being a Riemannian gradient and

  by orthogonality of $\psi(X)X$ and $\nabla \mathscr{N}(X^*)$

- For all $X_0 \in \mathbb{R}_*^{n\times p}$, the ω-limit points of $\varphi_t(X_0)$ belong to $\mathscr{C}$.
  Therefore, the landing system $\dot{X}(t) = -\Lambda\left(X(t)\right)$, converges to the set of critical points of f relative to $\mathrm{St}(p, n)$.

- For all $X_0 \in \mathbb{R}_*^{n\times p}$, if X^* is a local minimum and isolated critical point of f relative to $\mathrm{St}(p, n)$, and if X^* is an ω-limit point of $\varphi_t(X_0)$, then $\lim_{t\to\infty} \varphi_t(X_0) = X^*$.

Landing algorithm

- Discretize the flow
    - $X^{t+1} = X^t - \eta_t \Lambda\left(X^t\right),$ where $\psi(X)X + \lambda \nabla \mathcal{N}(X),$
- Numerical experiments with Stochastic gradient descent (SGD) to compare
    - Riemannian SGD (Retraction -QR)
    - ℓ_2-penalty (regularization with $+\frac{\lambda}{4}\|X^\top X - I_p\|_F^2$)
- Fixed step-size $\eta = 0.1$, landing parameter $\lambda = 1$

Online PCA

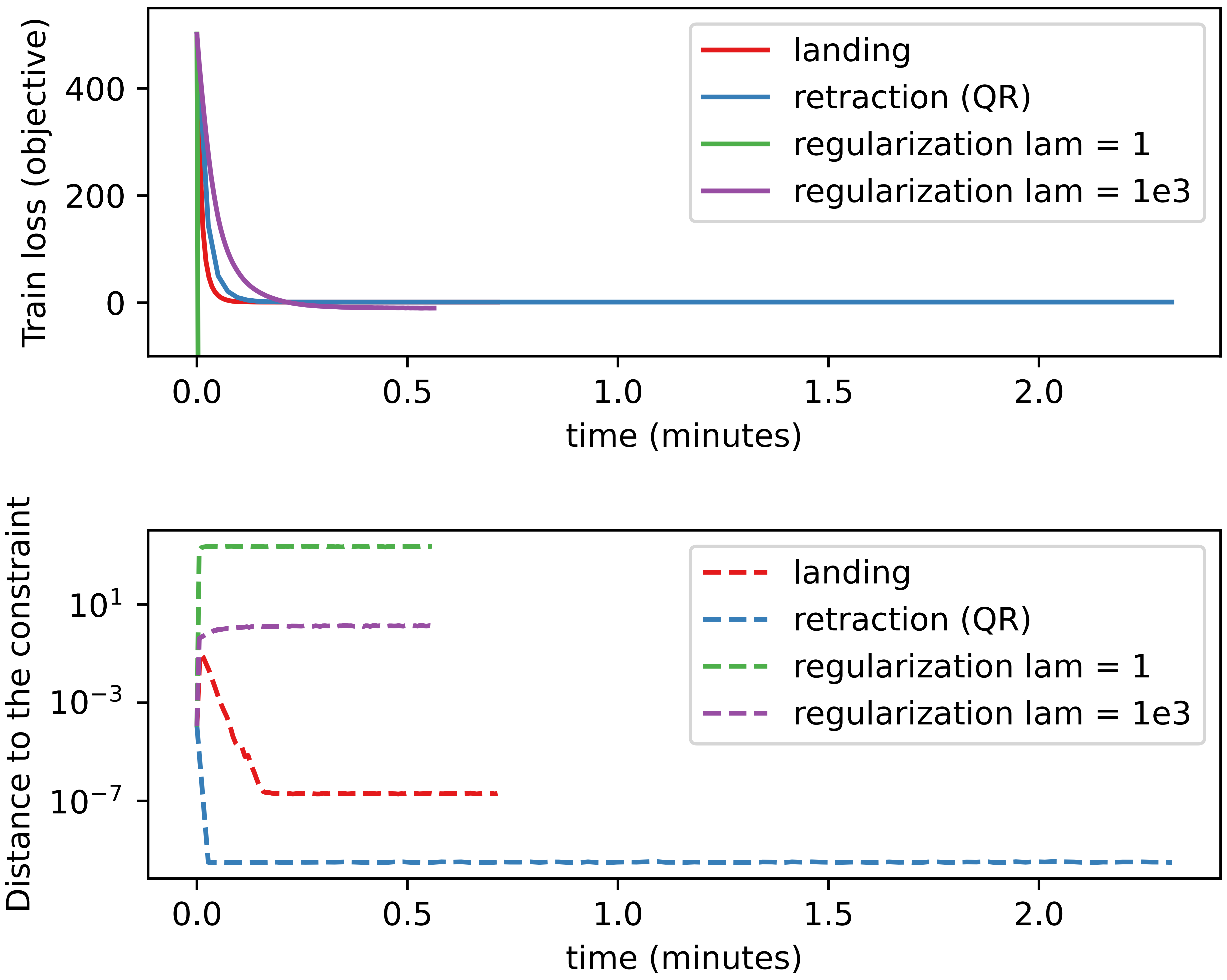
Consider

$$\min -\frac{1}{2}\|AX\|_F^2 \quad \text{s.t.} \quad X \in \mathrm{St}(p,n),$$

where

- $A \in \mathbb{R}^{m \times n}$, with $m = 10\,000, n = 2\,000$
- $p = 1\,000$

with Stochastic gradient descent with batch size of 128 rows.



Orthogonal convolutional neural network*

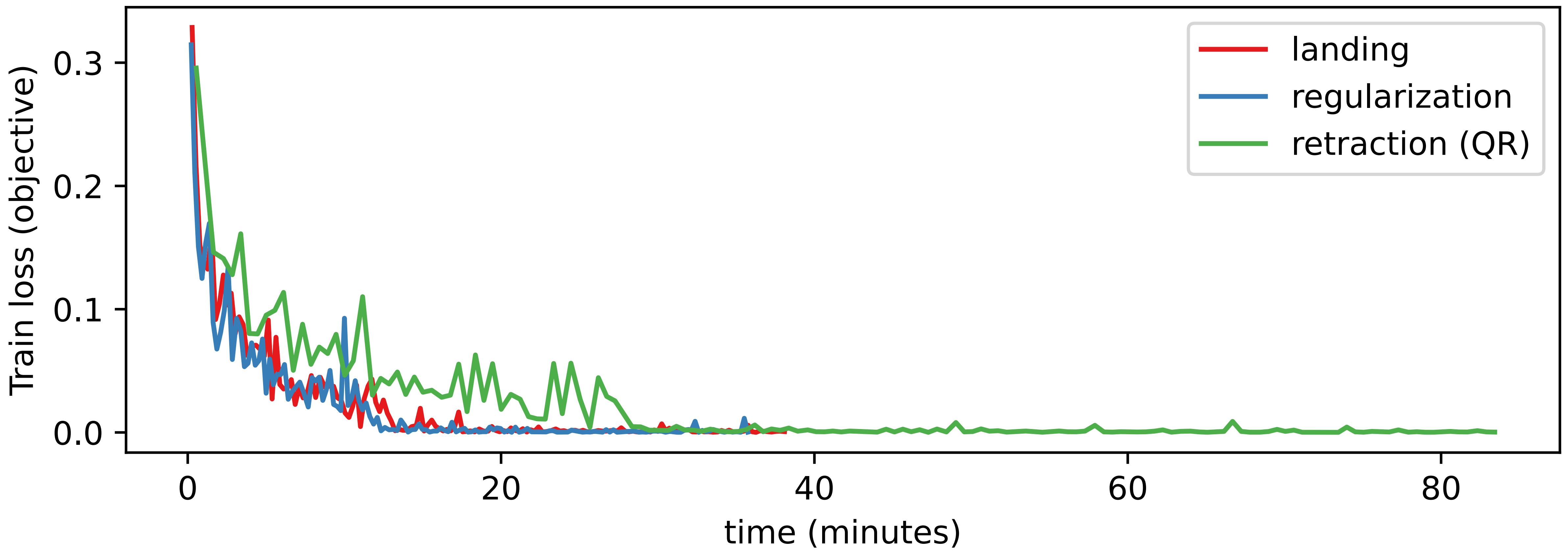
Consider

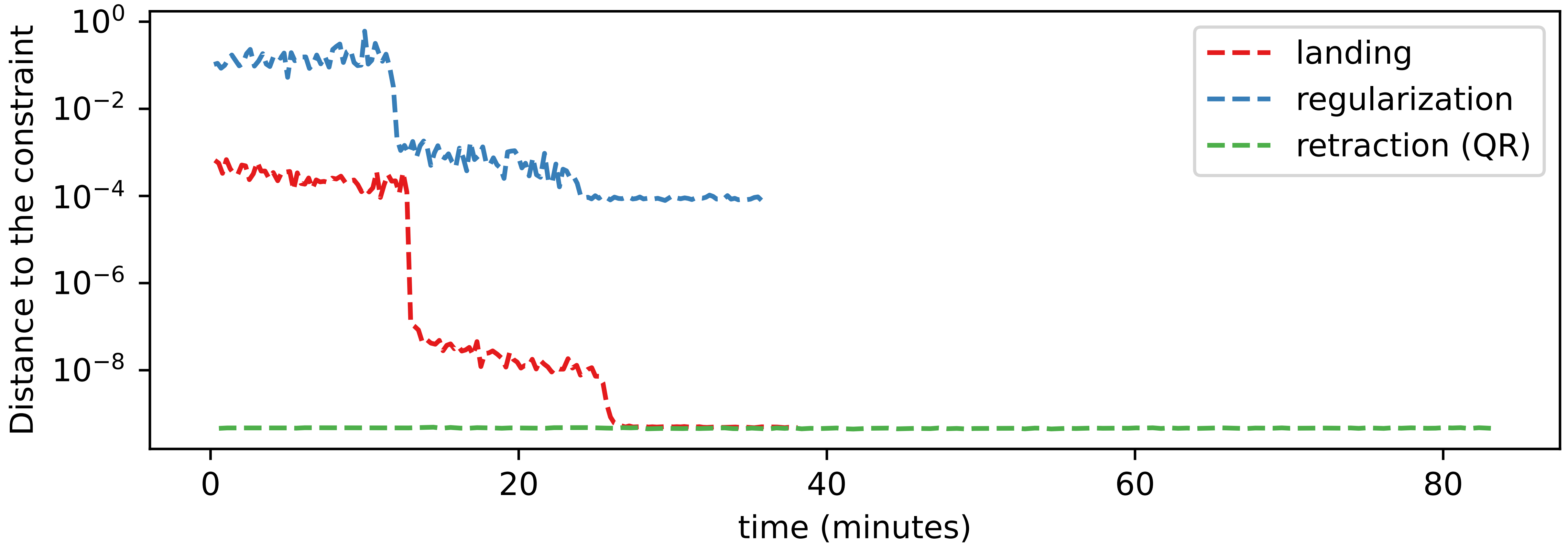
$$\min \sum_{i}^{N} \ell(f_\Theta(x_i), y_i) \quad \text{s.t.} \quad \forall \theta \in \Theta_{\text{orth}} : \theta_i \in \text{St}(p, n)$$

where

- $f_\Theta(\cdot)$ is a VGG16** convolutional neural network
- Θ_{orth} includes 13 matrices with size $\approx (1\,000)^2$
- trained on CIFAR-10

with a batch size of 128 samples, fixed step-size





* Orthogonal convolutional neural networks
Wang et al., CVPR 2020

** Very Deep Convolutional Net. for Large-Scale Image Recognition
Simonyan & Zisserman, ICLR 2015

Conclusions

- We propose a landing flow/algorithm and
    - provide a geometric interpretation of the field
    - analyze the continuous gradient flow and show its convergence to local minima
    - demonstrate with numerical experiments its efficiency
- Future work:
    - analysis of the discrete case
    - clever rules for the step-sizes in terms of λ and η
    - possible extensions, for example higher-order and acceleration