Convergence of Intrinsic Riemannian Proximal Gradient Methods

Hajg Jasa

joint work with

Ronny Bergmann, Paula John, and Max Pfeffer

MaGIC 2026 - Alta, Norway

April 7th, 2026

Riemannian Geometry: Notation

- Geodesic
  $\gamma_{pq} \colon [0,1] \to \mathcal{M}$ with initial velocity
  $\dot\gamma_{pq}(0) = X_p \in \mathcal{T}_p\mathcal{M}$
- Inner product
  $(\cdot,\cdot)_p : \mathcal{T}_p\mathcal{M} \times \mathcal{T}_p\mathcal{M} \to \mathbb{R}$
- Exponential map
  $\exp_p X_p = \gamma_{pq}(1) = q$
- Logarithmic map
  $\log_p q = \exp_p^{-1} q = X_p$
- Sectional curvature
  $\kappa \in [\kappa_{\mathrm{l}}, \kappa_{\mathrm{u}}]$

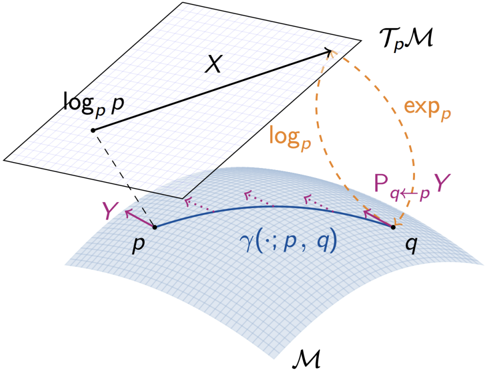


Figure: Courtesy of Ronny Bergmann.

Geodesic Convexity

- A set $\mathcal{C} \subseteq \mathcal{M}$ is called (strongly) geodesically convex if for all $p, q \in \mathcal{C}$ the geodesic $\gamma_{pq} \colon [0,1] \to \mathcal{M}$ (is unique and) lies in $\mathcal{C}$.

Geodesic Convexity

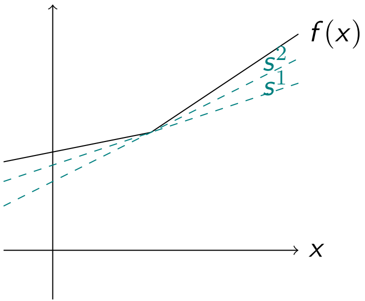
- A set $\mathcal{C} \subseteq \mathcal{M}$ is called (strongly) geodesically convex if for all $p, q \in \mathcal{C}$ the geodesic $\gamma_{pq} \colon [0,1] \to \mathcal{M}$ (is unique and) lies in $\mathcal{C}$.
- A function $f \colon \mathcal{C} \to \overline{\mathbb{R}}$ is called geodesically convex if for all $p, q \in \mathcal{C}$ the composition $f \circ \gamma_{pq}(t)$ is convex in the usual sense.

The Subdifferentials

For a convex function, the subdifferential is defined as

$$\partial f(x) = \left\{ s \in \mathbb{R}^n \,\middle|\, f(y) \geq f(x) + (s)^T (y - x) \text{ for all } y \in \mathbb{R}^n \right\}$$

and it is a non-empty, closed and convex subset.



The Subdifferentials

For a geodesically convex function, the subdifferential is defined as

$$\partial f(p) = \left\{ X_p \in \mathcal{T}_p\mathcal{M} \,\middle|\, f(q) \geq f(p) + \left(X_p, \log_p q \right) \text{ for all } q \in \mathcal{U} \right\}$$

with $\mathcal{U} \subseteq \mathcal{M}$ geodesically convex, and it is a non-empty, closed and convex subset.

The Subdifferentials

For a geodesically convex function, the subdifferential is defined as

$$\partial f(p) = \{X_p \in \mathcal{T}_p\mathcal{M} \,|\, f(q) \geq f(p) + (X_p, \log_p q) \text{ for all } q \in \mathcal{U}\}$$

with $\mathcal{U} \subseteq \mathcal{M}$ geodesically convex, and it is a non-empty, closed and convex subset.

For a locally Lipschitz function , define the generalized Clarke derivative

$$f^\circ(x; v) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + tv) - f(y)}{t}.$$

This can be used to define the *Clarke subdifferential*

$$\partial^\circ f(x) = \{s \in \mathbb{R}^n \,|\, (s\,,\, v) \leq f^\circ(x; v),\ \forall v \in \mathbb{R}^n\}.$$

which is nonempty, convex and compact.

The Subdifferentials

For a geodesically convex function, the subdifferential is defined as

$$\partial f(p) = \{X_p \in \mathcal{T}_p\mathcal{M} \,|\, f(q) \geq f(p) + (X_p, \log_p q) \text{ for all } q \in \mathcal{U}\}$$

with $\mathcal{U} \subseteq \mathcal{M}$ geodesically convex, and it is a non-empty, closed and convex subset.

For a locally Lipschitz function on a manifold, define the generalized Clarke derivative

$$f^\circ(p; Y_p) = (f \circ \exp_p)^\circ(0_p; Y_p).$$

This can be used to define the *Clarke subdifferential*

$$\partial^\circ f(p) = \{X_p \in \mathcal{T}_p\mathcal{M} \,|\, (X_p\,,\, Y_p) \leq f^\circ(0_p; Y_p),\ \forall Y_p \in \mathcal{T}_p\mathcal{M}\}.$$

which is nonempty, convex and compact.

The Proximal Operator

Given a function $f\colon \mathcal{M} \to \mathbb{R}$ and a real number $\lambda > 0$, the proximal operator of f is

$$\operatorname{prox}_{\lambda f}(p) := \operatorname*{arg\,min}_{q\in\mathcal{M}} f(q) + \frac{1}{2\lambda}\operatorname{dist}^2(q, p),$$

The Proximal Operator

Given a function $f\colon \mathcal{M} \to \mathbb{R}$ and a real number $\lambda > 0$, the proximal operator of f is

$$\operatorname{prox}_{\lambda f}(p) := \operatorname*{arg\,min}_{q \in \mathcal{M}} f(q) + \frac{1}{2\lambda}\operatorname{dist}^2(q, p),$$

If f is proper, lower semi-continuous, and geodesically convex, then $\operatorname{prox}_{\lambda f}$ is single-valued.

The Problem

Consider the following minimization problem

$$\text{minimize } f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

The Problem

Consider the following minimization problem

$$\operatorname{minimize} f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

where

- $\mathcal{M}$ is a Riemannian manifold

The Problem

Consider the following minimization problem

$$\text{minimize } f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

where

- $\mathcal{M}$ is a Riemannian manifold
  - complete
  - bounded sectional curvatures in $[\kappa_l, \kappa_u]$

The Problem

Consider the following minimization problem

$$\text{minimize } f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

where

- $\mathcal{M}$ is a Riemannian manifold
  - complete
  - bounded sectional curvatures in $[\kappa_l, \kappa_u]$
- and

The Problem

Consider the following minimization problem

$$\text{minimize } f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

where

- $\mathcal{M}$ is a Riemannian manifold
  - complete
  - bounded sectional curvatures in $[\kappa_l, \kappa_u]$
- and
  - $g\colon \mathcal{M} \to \overline{\mathbb{R}}$ is L_g-smooth
  - $h\colon \mathcal{M} \to \overline{\mathbb{R}}$ is (locally) Lipschitz continuous
  - f has compact level sets

The Problem

Consider the following minimization problem

$$\text{minimize } f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

where

- $\mathcal{M}$ is a Riemannian manifold
  - complete
  - bounded sectional curvatures in $[\kappa_l, \kappa_u]$
- and
  - $g\colon \mathcal{M} \to \overline{\mathbb{R}}$ is L_g-smooth
  - $h\colon \mathcal{M} \to \overline{\mathbb{R}}$ is (locally) Lipschitz continuous
  - f has compact level sets

Goal. Solve this problem with the proximal gradient method.

The Problem

Consider the following minimization problem

$$\text{minimize } f(p) = g(p) + h(p), \quad p \in \mathcal{M},$$

where

- $\mathcal{M}$ is a Hadamard manifold
  - complete
  - bounded sectional curvatures in $[\kappa_l, \kappa_u]$
- and
  - $g\colon \mathcal{M} \to \overline{\mathbb{R}}$ is proper, closed, L_g-smooth, and g-convex
  - $h\colon \mathcal{M} \to \overline{\mathbb{R}}$ is proper, closed, and g-convex
  - f has compact level sets

Goal. Solve this problem with the proximal gradient method.

Derivation of the Method

Main idea: Replace g by a proximal regularization of its linearization.

Derivation of the Method

Main idea: Replace g by a proximal regularization of its linearization.
For $\lambda > 0$ and a point $p \in \mathcal{M}$, compute a new candidate as

$$\operatorname*{arg\,min}_{q \in \mathcal{M}} g(p) + \left(\operatorname{grad} g(p)\,,\, \log_p q\right) + \frac{1}{2\lambda}\operatorname{dist}^2(p, q) + h(q).$$

Derivation of the Method

Main idea: Replace g by a proximal regularization of its linearization.
For $\lambda > 0$ and a point $p \in \mathcal{M}$, compute a new candidate as

$$\operatorname*{arg\,min}_{q \in \mathcal{M}} g(p) + \left(\operatorname{grad} g(p)\,,\ \log_p q\right) + \frac{1}{2\lambda} \operatorname{dist}^2(p, q) + h(q).$$

Taylor expand the distance term to second order and obtain

$$\operatorname*{arg\,min}_{q \in \mathcal{M}} g(p) + \frac{1}{2\lambda} \operatorname{dist}^2\left(\exp_p(-\lambda \operatorname{grad} g(p)), q\right) + h(q).$$

Derivation of the Method

Main idea: Replace g by a proximal regularization of its linearization. For $\lambda > 0$ and a point $p \in \mathcal{M}$, compute a new candidate as

$$\operatorname*{arg\,min}_{q \in \mathcal{M}} g(p) + \left(\operatorname{grad} g(p)\,,\, \log_p q\right) + \frac{1}{2\lambda}\operatorname{dist}^2(p,q) + h(q).$$

Taylor expand the distance term to second order and obtain

$$\operatorname*{arg\,min}_{q \in \mathcal{M}} g(p) + \frac{1}{2\lambda}\operatorname{dist}^2\left(\exp_p(-\lambda \operatorname{grad} g(p)), q\right) + h(q).$$

Since $g(p)$ does not change the minimizer, we get, approximately

$$\operatorname{prox}_{\lambda h}(\exp_p(-\lambda \operatorname{grad} g(p))) = \operatorname*{arg\,min}_{q \in \mathcal{M}} \frac{1}{2\lambda}\operatorname{dist}^2(\exp_p(-\lambda \operatorname{grad} g(p)), q) + h(q)$$

Some Definitions

The objective f is continuous and its level set is compact, hence it is lower bounded by $f_{opt} := \min_{q \in \mathcal{L}_{p(0)}} f(q)$.

Since h and $\operatorname{grad} g$ are continuous we obtain bounds $\alpha_1, \alpha_2, \alpha_g \in \mathbb{R}$ such that

$$\alpha_1 \leq h(q) \leq \alpha_2 \text{ for all } q \in \mathcal{L}_{p(0)}, \text{ and } \|\operatorname{grad} g(q)\| \leq \alpha_g \text{ for all } q \in \mathcal{L}_{p(0)}.$$

Some Definitions

The objective f is continuous and its level set is compact, hence it is lower bounded by $f_{opt} := \min_{q \in \mathcal{L}_{p(0)}} f(q)$.

Since h and $\operatorname{grad} g$ are continuous we obtain bounds $\alpha_1, \alpha_2, \alpha_g \in \mathbb{R}$ such that

$$\alpha_1 \leq h(q) \leq \alpha_2 \text{ for all } q \in \mathcal{L}_{p(0)}, \text{ and } \|\operatorname{grad} g(q)\| \leq \alpha_g \text{ for all } q \in \mathcal{L}_{p(0)}.$$

For real numbers κ_1, κ_2, s, define the quantities

$$\zeta_{1,\kappa_1}(s) := \begin{cases} 1 & \text{if } \kappa_1 \geq 0, \\ \sqrt{-\kappa_1}\, s \coth(\sqrt{-\kappa_1}\, s) & \text{if } \kappa_1 < 0, \end{cases}$$

$$\zeta_{2,\kappa_2}(s) := \begin{cases} 1 & \text{if } \kappa_2 \leq 0, \\ \sqrt{\kappa_2}\, s \cot(\sqrt{\kappa_2}\, s) & \text{if } \kappa_2 > 0, \end{cases}$$

$$\pi_\kappa := \begin{cases} \infty & \text{if } \kappa \leq 0, \\ \frac{\pi}{\sqrt{\kappa}} & \text{if } \kappa > 0, \end{cases}$$

important for taking curvature into account.

NTNU

Some Definitions II

For $p \in \mathcal{M}$ and $\lambda > 0$, the *iteration map* $\mathrm{T}_\lambda \colon \mathcal{M} \to \mathcal{M}$ that to each point p assigns a stationary point $\mathrm{T}_\lambda(p)$ of the function

$$H(\cdot, p, \lambda) \coloneqq h(\cdot) + \frac{1}{2\lambda} \operatorname{dist}^2(\cdot, \exp_p(-\lambda \operatorname{grad} g(p))),$$

and $\frac{1}{\lambda} \log_{\mathrm{T}_\lambda(q)} z(q) \in \partial h(\mathrm{T}_\lambda(q))$, with $z(p) = \exp_p(-\lambda \operatorname{grad} g(p))$.

Some Definitions II

For $p \in \mathcal{M}$ and $\lambda > 0$, the *iteration map* $\mathrm{T}_\lambda \colon \mathcal{M} \to \mathcal{M}$ that to each point p assigns a stationary point $\mathrm{T}_\lambda(p)$ of the function

$$H(\cdot, p, \lambda) \coloneqq h(\cdot) + \frac{1}{2\lambda} \operatorname{dist}^2(\cdot, \exp_p(-\lambda \operatorname{grad} g(p))),$$

and $\frac{1}{\lambda} \log_{\mathrm{T}_\lambda(q)} z(q) \in \partial h(\mathrm{T}_\lambda(q))$, with $z(p) = \exp_p(-\lambda \operatorname{grad} g(p))$.

If h is e. g. geodesically convex,

$$\mathrm{T}_\lambda(p) = \operatorname{prox}_{\lambda h}\Bigl(\exp_p(-\lambda \operatorname{grad} g(p))\Bigr) \in \mathcal{M}.$$

NTNU

Some Definitions II

For $p \in \mathcal{M}$ and $\lambda > 0$, the *iteration map* $\mathrm{T}_\lambda \colon \mathcal{M} \to \mathcal{M}$ that to each point p assigns a stationary point $\mathrm{T}_\lambda(p)$ of the function

$$H(\cdot, p, \lambda) := h(\cdot) + \frac{1}{2\lambda} \operatorname{dist}^2(\cdot, \exp_p(-\lambda \operatorname{grad} g(p))),$$

and $\frac{1}{\lambda} \log_{\mathrm{T}_\lambda(q)} z(q) \in \partial h(\mathrm{T}_\lambda(q))$, with $z(p) = \exp_p(-\lambda \operatorname{grad} g(p))$.

If h is e. g. geodesically convex,

$$\mathrm{T}_\lambda(p) = \operatorname{prox}_{\lambda h}\Bigl(\exp_p(-\lambda \operatorname{grad} g(p))\Bigr) \in \mathcal{M}.$$

The nontrivial task is analyzing stationarity/optimality due to the nonlinearity of $\log_p \mathrm{T}_\lambda(p)$.

NTNU

Some Definitions II

For $p \in \mathcal{M}$ and $\lambda > 0$, the *iteration map* $\mathrm{T}_\lambda \colon \mathcal{M} \to \mathcal{M}$ that to each point p assigns a stationary point $\mathrm{T}_\lambda(p)$ of the function

$$H(\cdot, p, \lambda) := h(\cdot) + \frac{1}{2\lambda} \operatorname{dist}^2(\cdot, \exp_p(-\lambda \operatorname{grad} g(p))),$$

and $\frac{1}{\lambda} \log_{\mathrm{T}_\lambda(q)} z(q) \in \partial h(\mathrm{T}_\lambda(q))$, with $z(p) = \exp_p(-\lambda \operatorname{grad} g(p))$.

If h is e. g. geodesically convex,

$$\mathrm{T}_\lambda(p) = \operatorname{prox}_{\lambda h}\Bigl(\exp_p(-\lambda \operatorname{grad} g(p))\Bigr) \in \mathcal{M}.$$

The nontrivial task is analyzing stationarity/optimality due to the nonlinearity of $\log_p \mathrm{T}_\lambda(p)$.

This is overcome via Riemannian cosine inequalities:

$$2(\log_q y\,,\ \log_q p) \leq \zeta_{1,\kappa_{\mathrm{l}}}(s) \operatorname{dist}^2(q, y) + \operatorname{dist}^2(p, q) - \operatorname{dist}^2(p, y),$$
$$2(\log_q y\,,\ \log_q p) \geq \zeta_{2,\kappa_{\mathrm{u}}}(s) \operatorname{dist}^2(q, y) + \operatorname{dist}^2(p, q) - \operatorname{dist}^2(p, y).$$

Technical Lemma

Let $p \in \mathcal{L}_{p^{(0)}}$ and $\delta > 0$,

$$\lambda_\delta := \frac{\sqrt{4\,(\alpha_2 - \alpha_1)^2 + \frac{{\pi_{\kappa_\mathrm{u}}}^2}{(2+\delta)^2}\,{\alpha_g}^2} - 2\,(\alpha_2 - \alpha_1)}{2\,{\alpha_g}^2},$$

and

$$\zeta_\delta := \zeta_{2,\kappa_\mathrm{u}}\left(\frac{\pi_{\kappa_\mathrm{u}}}{2+\delta}\right) = \begin{cases} 1 & \text{if } \kappa_\mathrm{u} \leq 0, \\ \frac{\pi}{(2+\delta)\sqrt{\kappa_\mathrm{u}}} \cot\left(\frac{\pi}{(2+\delta)\sqrt{\kappa_\mathrm{u}}}\right) & \text{if } \kappa_\mathrm{u} > 0. \end{cases}$$

Then, for all $\lambda \in (0, \lambda_\delta]$ and $z(p) = \exp_p(-\lambda \operatorname{grad} g(p))$,

$$\operatorname{dist}(p, z(p)) + \operatorname{dist}(\mathrm{T}_\lambda(p), z(p)) \leq \frac{\pi_{\kappa_\mathrm{u}}}{2+\delta},$$

and

$$\zeta_{2,\kappa_\mathrm{u}}(\operatorname{dist}(p, z(p)) + \operatorname{dist}(\mathrm{T}_\lambda(p), z(p))) \geq \zeta_\delta > 0.$$

NCRPG: Nonconvex Case

Data: g, $\operatorname{grad} g$, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

end

NCRPG: Nonconvex Case

Data: g, $\operatorname{grad} g$, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

  find a stationary point $p^{(k+1)}$ of $H(\cdot, p^{(k)}, \lambda^{(k)})$

end

NTNU

NCRPG: Nonconvex Case

Data: g, $\operatorname{grad} g$, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

    find a stationary point $p^{(k+1)}$ of $H(\cdot, p^{(k)}, \lambda^{(k)})$

    such that $H(p^{(k+1)}, p^{(k)}, \lambda^{(k)}) \leq H(p^{(k)}, p^{(k)}, \lambda^{(k)})$

end

NCRPG: Nonconvex Case

Data: g, $\operatorname{grad} g$, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

  find a stationary point $p^{(k+1)}$ of $H(\cdot, p^{(k)}, \lambda^{(k)})$

  such that $H(p^{(k+1)}, p^{(k)}, \lambda^{(k)}) \leq H(p^{(k)}, p^{(k)}, \lambda^{(k)})$

  Set $k := k + 1$

end

NCRPG: Nonconvex Case

Data: g, $\operatorname{grad} g$, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

  find a stationary point $p^{(k+1)}$ of $H(\cdot, p^{(k)}, \lambda^{(k)})$

  such that $H(p^{(k+1)}, p^{(k)}, \lambda^{(k)}) \leq H(p^{(k)}, p^{(k)}, \lambda^{(k)})$

  Set $k := k + 1$

end

Result: $p^{(k_*)}$, for some $k_* \in \mathbb{N}$.

NCRPG: Nonconvex Case

Data: g, grad g, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

  find a stationary point $p^{(k+1)}$ of $H(\cdot, p^{(k)}, \lambda^{(k)})$

  such that $H(p^{(k+1)}, p^{(k)}, \lambda^{(k)}) \leq H(p^{(k)}, p^{(k)}, \lambda^{(k)})$

  Set $k := k + 1$

end

Result: $p^{(k_*)}$, for some $k_* \in \mathbb{N}$.

Backtracking strategy

Data: f, $p^{(k)}$, initial guess $s > 0$, $\eta \in (0, 1)$ and $\beta \in (0, \frac{\zeta_\delta}{2})$.

Set $\lambda^{(k)} = s$

NCRPG: Nonconvex Case

Data: g, $\operatorname{grad} g$, h, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

  find a stationary point $p^{(k+1)}$ of $H(\cdot, p^{(k)}, \lambda^{(k)})$

  such that $H(p^{(k+1)}, p^{(k)}, \lambda^{(k)}) \leq H(p^{(k)}, p^{(k)}, \lambda^{(k)})$

  Set $k := k + 1$

end

Result: $p^{(k_*)}$, for some $k_* \in \mathbb{N}$.

Backtracking strategy

Data: f, $p^{(k)}$, initial guess $s > 0$, $\eta \in (0, 1)$ and $\beta \in (0, \frac{\zeta_\delta}{2})$.

Set $\lambda^{(k)} = s$

while $f(p^{(k)}) - f\left(\mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)\right) < \frac{\beta}{\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)) $ **do**

  Set $\lambda^{(k)} = \eta\lambda^{(k)}$

end

Convergence: Nonconvex Case

NTNU

Given either a constant step-size $\lambda^{(k)} := \lambda \in \left(0, \min\left\{\lambda_\delta, \frac{\zeta_\delta}{L_g}\right\}\right)$ or one that is chosen by the backtracking strategy, then

1. the sequence $\left(f(p^{(k)})\right)$ is nonincreasing, and $f(p^{(k+1)}) < f(p^{(k)})$ if and only if $p^{(k)}$ is not a stationary point of the original problem;
2. $\lambda^{(k)} \operatorname{dist}(p^{(k)}, p^{(k+1)}) \to 0$ as $k \to \infty$;
3. $\min_{n=0,1,\ldots,k} \lambda^{(n)} \operatorname{dist}(p^{(n)}, p^{(n+1)}) = O\left(\frac{1}{\sqrt{k+1}}\right)$;
4. all limit points of the sequence $(p^{(k)})$ are stationary points of f;
5. the algorithm returns an ε-stationary point in $\mathcal{O}\left(\frac{\omega}{\varepsilon^2}\right)$, where ω includes terms that depend on the curvature bounds of $\mathcal{M}$.

[Bergmann, HJ, et al. 2025b, Lemma 4.7, Theorem 4.9, Theorem 4.10]

CRPG: Convex Case

Data: g, $\operatorname{grad} g$, h, $\operatorname{prox}_{\lambda h}$, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

end

NTNU

CRPG: Convex Case

Data: g, $\operatorname{grad} g$, h, $\operatorname{prox}_{\lambda h}$, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

$$p^{(k+1)} = \operatorname{prox}_{\lambda^{(k)} h}\Big(\exp_{p^{(k)}}\big(-\lambda^{(k)} \operatorname{grad} g(p^{(k)})\big)\Big)$$

end

CRPG: Convex Case

Data: g, $\operatorname{grad} g$, h, $\operatorname{prox}_{\lambda h}$, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

$$p^{(k+1)} = \operatorname{prox}_{\lambda^{(k)} h}\Bigl(\exp_{p^{(k)}}\bigl(-\lambda^{(k)} \operatorname{grad} g(p^{(k)})\bigr)\Bigr)$$

Set $k := k+1$

end

CRPG: Convex Case

Data: g, $\operatorname{grad} g$, h, $\operatorname{prox}_{\lambda h}$, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

$$p^{(k+1)} = \operatorname{prox}_{\lambda^{(k)} h}\Bigl(\exp_{p^{(k)}}\bigl(-\lambda^{(k)} \operatorname{grad} g(p^{(k)})\bigr)\Bigr)$$

Set $k := k+1$

end

Result: $p^{(k_*)}$, for some $k_* \in \mathbb{N}$.

CRPG: Convex Case

Data: g, $\operatorname{grad} g$, h, $\operatorname{prox}_{\lambda h}$, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

$$p^{(k+1)} = \operatorname{prox}_{\lambda^{(k)} h}\Bigl(\exp_{p^{(k)}}\bigl(-\lambda^{(k)} \operatorname{grad} g(p^{(k)})\bigr)\Bigr)$$

Set $k := k + 1$

end

Result: $p^{(k_*)}$, for some $k_* \in \mathbb{N}$.

Backtracking strategy

Data: f, $p^{(k)}$, initial guess $s \in \left(0, \frac{2}{L_g}\right)$, $\eta \in (0,1)$ and $\theta \geq 1$.

Set $\lambda^{(-1)} = s$, and $\lambda^{(k)} = \min\{s, \theta\lambda^{(k-1)}\}$

CRPG: Convex Case

Data: g, $\operatorname{grad} g$, h, $\operatorname{prox}_{\lambda h}$, a sequence $\lambda^{(k)}$, an initial point $p^{(0)} \in \mathcal{M}$.

while *convergence criterion is not fulfilled* **do**

$$p^{(k+1)} = \operatorname{prox}_{\lambda^{(k)} h}\Bigl(\exp_{p^{(k)}}\bigl(-\lambda^{(k)} \operatorname{grad} g(p^{(k)})\bigr)\Bigr)$$

Set $k := k + 1$

end

Result: $p^{(k_*)}$, for some $k_* \in \mathbb{N}$.

Backtracking strategy

Data: f, $p^{(k)}$, initial guess $s \in \left(0, \frac{2}{L_g}\right)$, $\eta \in (0, 1)$ and $\theta \geq 1$.

Set $\lambda^{(-1)} = s$, and $\lambda^{(k)} = \min\{s, \theta\lambda^{(k-1)}\}$

while $g\left(\mathrm{T}_{\lambda^{(k)}}(p^{(k)})\right) >$
$g(p^{(k)}) + \left(\operatorname{grad} g(p^{(k)}), \mathrm{T}_{\lambda^{(k)}}(p^{(k)})\right) + \frac{1}{\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, \mathrm{T}_{\lambda^{(k)}}(p^{(k)}))$ **do**

Set $\lambda^{(k)} = \eta\lambda^{(k)}$

end

Convergence: Fundamental Prox-Grad Inequality

In Euclidean space, L_g-smoothness of g and convexity of h are enough to establish

$$f(x) - f(\mathrm{T}_\lambda(y)) \geq l_g(x, y) + \frac{1}{2\lambda}\|x - \mathrm{T}_\lambda(y)\|^2 - \frac{1}{2\lambda}\|x - y\|^2,$$

with

$$l_g(x, y) = g(x) - g(y) - \nabla g(y)^\top (x - y).$$

Convergence: Fundamental Prox-Grad Inequality

In Euclidean space, L_g-smoothness of g and convexity of h are enough to establish

$$f(x) - f(\mathrm{T}_\lambda(y)) \geq l_g(x, y) + \frac{1}{2\lambda}\|x - \mathrm{T}_\lambda(y)\|^2 - \frac{1}{2\lambda}\|x - y\|^2,$$

with

$$l_g(x, y) = g(x) - g(y) - \nabla g(y)^\top (x - y).$$

When g is also convex, by telescoping one easily gets a convergence rate for the Proximal Gradient method:

$$f(x^{(k)}) - f_{\text{opt}} \leq \frac{\alpha L_g \|x^{(0)} - x^*\|^2}{2k},$$

where α is a bound on the step-size.

Fundamental Prox-Grad Inequality II

On a uniquely geodesic Riemannian manifold, with a different strategy we get

$$
\begin{aligned}
f(p) - f(\mathrm{T}_\lambda(q)) \geq l_g(p,q) &+ \frac{1+\zeta_{2,\kappa_\mathrm{u}}(D_3)}{4\lambda}\operatorname{dist}^2(p,\mathrm{T}_\lambda(q)) - \frac{1+\zeta_{1,\kappa_\mathrm{l}}(D_2)}{4\lambda}\operatorname{dist}^2(p,q)\\
&+ \frac{\zeta_{2,\kappa_\mathrm{u}}(D_3)-1}{4\lambda}\operatorname{dist}^2(z(q),\mathrm{T}_\lambda(q)) + \frac{\zeta_{2,\kappa_\mathrm{u}}(D_1)-1}{4\lambda}\operatorname{dist}^2(q,\mathrm{T}_\lambda(q))\\
&+ \frac{\zeta_{2,\kappa_\mathrm{u}}(D_1)-\zeta_{1,\kappa_\mathrm{l}}(D_2)}{4\lambda}\operatorname{dist}^2(q,z(q)),
\end{aligned}
$$

Fundamental Prox-Grad Inequality II

On a uniquely geodesic Riemannian manifold, with a different strategy we get

$$
\begin{aligned}
f(p) - f(\mathrm{T}_\lambda(q)) \geq l_g(p,q) &+ \frac{1+\zeta_{2,\kappa_\mathrm{u}}(D_3)}{4\lambda}\operatorname{dist}^2(p, \mathrm{T}_\lambda(q)) - \frac{1+\zeta_{1,\kappa_\mathrm{l}}(D_2)}{4\lambda}\operatorname{dist}^2(p,q)\\
&+ \frac{\zeta_{2,\kappa_\mathrm{u}}(D_3)-1}{4\lambda}\operatorname{dist}^2(z(q), \mathrm{T}_\lambda(q)) + \frac{\zeta_{2,\kappa_\mathrm{u}}(D_1)-1}{4\lambda}\operatorname{dist}^2(q, \mathrm{T}_\lambda(q))\\
&+ \frac{\zeta_{2,\kappa_\mathrm{u}}(D_1)-\zeta_{1,\kappa_\mathrm{l}}(D_2)}{4\lambda}\operatorname{dist}^2(q, z(q)),
\end{aligned}
$$

with $D_1 = D(q, z(q), \mathrm{T}_\lambda(q))$, $D_2 = D(q, z(q), p)$, and $D_3 = D(\mathrm{T}_\lambda(q), z(q), p)$, where $D(p,q,r)$ is the diameter of the uniquely geodesic triangle with vertices $p, q, r \in \mathcal{M}$, and $z(q) := \exp_q(-\lambda \operatorname{grad} g(q))$.

Fundamental Prox-Grad Inequality III

On a Hadamard manifold, $\kappa_{\mathrm{u}} = 0$, implying $\zeta_2 \equiv 1$, and assuming g to be geodesically convex yields

$$f(p) - f(\mathrm{T}_\lambda(q)) \geq \frac{1}{2\lambda}\operatorname{dist}^2(p, \mathrm{T}_\lambda(q)) - \frac{1 + \zeta_{1,\kappa_{\mathrm{l}}}(\operatorname{dist}(q, z(q)))}{4\lambda}\operatorname{dist}^2(p, q)$$
$$+ \frac{1 - \zeta_{1,\kappa_{\mathrm{l}}}(\operatorname{dist}(p, q))}{4\lambda}\operatorname{dist}^2(q, z(q)).$$

Fundamental Prox-Grad Inequality III

On a Hadamard manifold, $\kappa_{\mathrm{u}} = 0$, implying $\zeta_2 \equiv 1$, and assuming g to be geodesically convex yields

$$f(p) - f(\mathrm{T}_\lambda(q)) \geq \frac{1}{2\lambda}\operatorname{dist}^2(p, \mathrm{T}_\lambda(q)) - \frac{1 + \zeta_{1,\kappa_1}(\operatorname{dist}(q, z(q)))}{4\lambda}\operatorname{dist}^2(p, q) + \frac{1 - \zeta_{1,\kappa_1}(\operatorname{dist}(p, q))}{4\lambda}\operatorname{dist}^2(q, z(q)).$$

Let $R := \operatorname{dist}(p^{(0)}, p^*)$. If $g \equiv 0$, then $z(q) = q$ and by telescoping we obtain a convergence rate for the "pure" Proximal Point method with a bound β on the step-size:

$$f(p^{(k)}) - f_{\mathrm{opt}} \leq \frac{\beta L_g R^2}{2k}.$$

NTNU

Fundamental Prox-Grad Inequality III

On a Hadamard manifold, $\kappa_{\mathrm{u}} = 0$, implying $\zeta_2 \equiv 1$, and assuming g to be geodesically convex yields

$$f(p) - f(\mathrm{T}_\lambda(q)) \geq \frac{1}{2\lambda}\operatorname{dist}^2(p, \mathrm{T}_\lambda(q)) - \frac{1 + \zeta_{1,\kappa_1}(\operatorname{dist}(q, z(q)))}{4\lambda}\operatorname{dist}^2(p, q) + \frac{1 - \zeta_{1,\kappa_1}(\operatorname{dist}(p, q))}{4\lambda}\operatorname{dist}^2(q, z(q)).$$

Let $R := \operatorname{dist}(p^{(0)}, p^*)$. If $g \equiv 0$, then $z(q) = q$ and by telescoping we obtain a convergence rate for the “pure” Proximal Point method with a bound β on the step-size:

$$f(p^{(k)}) - f_{\mathrm{opt}} \leq \frac{\beta L_g R^2}{2k}.$$

If $h \equiv 0$, then $z(q) = \mathrm{T}_\lambda(q)$, and with a bit more work we get a rate for the “pure” Gradient Descent method with a constant step-size λ:

$$f(p^{(k)}) - f_{\mathrm{opt}} \leq \frac{R^2 + 2\lambda\zeta_{1,\kappa_1}(R)(f(p^{(0)}) - f_{\mathrm{opt}})}{2\lambda(\zeta_{1,\kappa_1}(R) + k)}.$$

Fundamental Prox-Grad Inequality IV

But if in general, $g, h \neq 0$, we don't get any nice telescoping: the curvature terms mix the gradient and proximal updates without cancellations!

Fundamental Prox-Grad Inequality IV

But if in general, $g, h \neq 0$, we don't get any nice telescoping: the curvature terms mix the gradient and proximal updates without cancellations!
One would hope to at least establish quasi-Féjer monotonicity, but we only get

$$\mathrm{dist}^2(p^{(k+1)}, p^*) \leq \frac{\zeta_{1,\kappa_1}(\mathrm{dist}(p^{(k)}, z(p^{(k)}))) + 1}{2}\,\mathrm{dist}^2(p^{(k)}, p^*) + \frac{\zeta_{1,\kappa_1}(\mathrm{dist}(p^*, p^{(k)})) - 1}{2}\,\mathrm{dist}^2(p^{(k)}, z(p^{(k)})),$$

and since in general $\mathrm{dist}(p^{(k)}, z(p^{(k)})) \not\to 0$, no such monotonicity can be shown.

Convergence: Convex Case

Given either a constant step-size $\lambda^{(k)} = \frac{1}{L_g}$ or one that is chosen by the backtracking strategy, then via the cosine laws, we show the algorithm returns an ε-stationary point in $\mathcal{O}\left(\frac{\Omega}{\varepsilon}\right)$, where Ω includes terms that depend on the curvature bounds of $\mathcal{M}$,

Convergence: Convex Case

Given either a constant step-size $\lambda^{(k)} = \frac{1}{L_g}$ or one that is chosen by the backtracking strategy, then via the cosine laws, we show the algorithm returns an ε-stationary point in $\mathcal{O}\left(\frac{\Omega}{\varepsilon}\right)$, where Ω includes terms that depend on the curvature bounds of $\mathcal{M}$, and there are regimes where the optimal gap $\Delta^{(k)} := f(p^{(k)}) - f_{\text{opt}}$ may decrease linearly:

Convergence: Convex Case

Given either a constant step-size $\lambda^{(k)} = \frac{1}{L_g}$ or one that is chosen by the backtracking strategy, then via the cosine laws, we show the algorithm returns an ε-stationary point in $\mathcal{O}\left(\frac{\Omega}{\varepsilon}\right)$, where Ω includes terms that depend on the curvature bounds of $\mathcal{M}$, and there are regimes where the optimal gap $\Delta^{(k)} := f(p^{(k)}) - f_{\text{opt}}$ may decrease linearly:

- if at iteration $k-1$, it holds that $\frac{\lambda^{(k-1)}\Delta^{(k-1)}}{\zeta_{1,\kappa_1}(D^{(k-1)})\,\text{dist}^2(p^{(k-1)},p^*)} \geq 1$, then

$$\Delta^{(k)} \leq \frac{1}{2}\Delta^{(k-1)};$$

Convergence: Convex Case

Given either a constant step-size $\lambda^{(k)} = \frac{1}{L_g}$ or one that is chosen by the backtracking strategy, then via the cosine laws, we show the algorithm returns an ε-stationary point in $\mathcal{O}\left(\frac{\Omega}{\varepsilon}\right)$, where Ω includes terms that depend on the curvature bounds of $\mathcal{M}$, and there are regimes where the optimal gap $\Delta^{(k)} := f(p^{(k)}) - f_{\text{opt}}$ may decrease linearly:

- if at iteration $k-1$, it holds that $\frac{\lambda^{(k-1)}\Delta^{(k-1)}}{\zeta_{1,\kappa_1}(D^{(k-1)})\,\text{dist}^2(p^{(k-1)},p^*)} \geq 1$, then

$$\Delta^{(k)} \leq \frac{1}{2}\Delta^{(k-1)};$$

- otherwise,

$$\Delta^{(k)} \leq \left(1 - \frac{\beta\Delta^{(k-1)}}{2L_g\zeta_{1,\kappa_1}(R_\alpha)R^2}\right)\Delta^{(k-1)},$$

NTNU

Convergence: Convex Case

Given either a constant step-size $\lambda^{(k)} = \frac{1}{L_g}$ or one that is chosen by the backtracking strategy, then via the cosine laws, we show the algorithm returns an ε-stationary point in $\mathcal{O}\left(\frac{\Omega}{\varepsilon}\right)$, where Ω includes terms that depend on the curvature bounds of $\mathcal{M}$, and there are regimes where the optimal gap $\Delta^{(k)} := f(p^{(k)}) - f_{\mathrm{opt}}$ may decrease linearly:

- if at iteration $k-1$, it holds that $\frac{\lambda^{(k-1)}\Delta^{(k-1)}}{\zeta_{1,\kappa_1}(D^{(k-1)})\,\mathrm{dist}^2(p^{(k-1)},p^*)} \geq 1$, then

$$\Delta^{(k)} \leq \frac{1}{2}\Delta^{(k-1)};$$

- otherwise,

$$\Delta^{(k)} \leq \left(1 - \frac{\beta\Delta^{(k-1)}}{2L_g\zeta_{1,\kappa_1}(R_\alpha)R^2}\right)\Delta^{(k-1)},$$

with $R := \mathrm{diam}(\mathcal{L}_{p^{(0)}})$, $D^{(n)} := \mathrm{dist}(p^{(n)}, z(p^{(n)}))$, $R_\alpha \geq \mathrm{dist}(p^{(n)}, z(p^{(n)}))$, $\alpha = \beta = 1$ if a constant stepsize is chosen, and $\alpha = sL_g$, and $\beta = \min\{sL_g, \eta\}$ for some $\eta \in (0,1)$ if the backtracking procedure is used.

Convex Case with Positive Curvature

Two challenges:

- convexity of g, h cannot be guaranteed globally;
- one of the cosine laws introduces a curvature terms that needs to be handled carefully:

$$-\frac{\zeta_{2,\kappa_{\mathrm{u}}}(\operatorname{diam}(p^{(k)}, z(p^{(k)}), p^{(k+1)}))}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, p^{(k+1)}).$$

Convex Case with Positive Curvature

Two challenges:

- convexity of g, h cannot be guaranteed globally;
- one of the cosine laws introduces a curvature terms that needs to be handled carefully:

$$-\frac{\zeta_{2,\kappa_\mathrm{u}}(\operatorname{diam}(p^{(k)}, z(p^{(k)}), p^{(k+1)}))}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, p^{(k+1)}).$$

Adapting the step-size selection may save the day: choose $\lambda^{(k)} = \frac{\zeta_\delta}{L_g}$ in the constant case, otherwise backtrack until

$$g\left(\mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)\right) \leq g(p^{(k)}) + \left(\operatorname{grad} g(p^{(k)}),\ \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)\right) + \frac{\zeta_\delta}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)),$$

which terminates finitely because g is L_g-smooth.

Convex Case with Positive Curvature

Two challenges:

- convexity of g, h cannot be guaranteed globally;
- one of the cosine laws introduces a curvature terms that needs to be handled carefully:

$$-\frac{\zeta_{2,\kappa_{\mathrm{u}}}(\operatorname{diam}(p^{(k)}, z(p^{(k)}), p^{(k+1)}))}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, p^{(k+1)}).$$

Adapting the step-size selection may save the day: choose $\lambda^{(k)} = \frac{\zeta_\delta}{L_g}$ in the constant case, otherwise backtrack until

$$g\left(\mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)\right) \leq g(p^{(k)}) + \left(\operatorname{grad} g(p^{(k)})\,,\, \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)\right) + \frac{\zeta_\delta}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)),$$

which terminates finitely because g is L_g-smooth. This should be enough to adapt the previous proof since $\zeta_\delta \leq \zeta_{2,\kappa_{\mathrm{u}}}(\operatorname{diam}(p^{(k)}, z(p^{(k)}), p^{(k+1)}))$ for all k, and the new term is compensated for.

Convex Case with Positive Curvature

Two challenges:

- convexity of g, h cannot be guaranteed globally;
- one of the cosine laws introduces a curvature terms that needs to be handled carefully:

$$-\frac{\zeta_{2,\kappa_{\mathrm{u}}}(\operatorname{diam}(p^{(k)}, z(p^{(k)}), p^{(k+1)}))}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, p^{(k+1)}).$$

Adapting the step-size selection may save the day: choose $\lambda^{(k)} = \frac{\zeta_\delta}{L_g}$ in the constant case, otherwise backtrack until

$$g\left(\mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)\right) \leq g(p^{(k)}) + (\operatorname{grad} g(p^{(k)}), \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)) + \frac{\zeta_\delta}{2\lambda^{(k)}} \operatorname{dist}^2(p^{(k)}, \mathrm{T}_{\lambda^{(k)}}\left(p^{(k)}\right)),$$

which terminates finitely because g is L_g-smooth. This should be enough to adapt the previous proof since $\zeta_\delta \leq \zeta_{2,\kappa_{\mathrm{u}}}(\operatorname{diam}(p^{(k)}, z(p^{(k)}), p^{(k+1)}))$ for all k, and the new term is compensated for. This also makes the backtracking independent of L_g!

Implementation

The algorithm is implemented in Julia using `Manopt.jl` ([Bergmann 2022]) and `Manifolds.jl` ([Axen et al. 2023]). A solver call looks like

```
p_star = proximal_gradient_method(M, f, g, grad_g, p0;
          prox_nonsmooth=prox_λh)
```

where

- `M` is a Riemannian manifold
- `f` is the objective function
- `g` is the smooth part of f
- `grad_g` is the gradient of g
- `p0` is an initial point on the manifold
- `prox_nonsmooth` is the proximal operator of h

The default stopping criterion for the algorithm is set to

$$\operatorname{dist}(p^{(k-1)}, p^{(k)}) \leq 10^{-8}.$$

Conclusion and Future Work

In summary:

- introduced an intrinsic Riemannian Proximal Gradient Method
- discussed convergence rates

To do:

- finish convex adaptation to manifolds with $\kappa_l \geq 0$

NTNU

Thank you for your attention!

Selected References

- Axen, S. D., M. Baran, R. Bergmann, and K. Rzecki (2023). "Manifolds.jl: An Extensible Julia Framework for Data Analysis on Manifolds". In: *ACM Transactions on Mathematical Software* 49.4. DOI: 10.1145/3618296. arXiv: 2106.08777.
- Bergmann, Ronny (2022). "Manopt.jl: Optimization on Manifolds in Julia". In: *Journal of Open Source Software* 7.70, p. 3866. DOI: 10.21105/joss.03866.
- Bergmann, Ronny, HJ, Paula John, and Max Pfeffer (2025a). *The Intrinsic Riemannian Proximal Gradient Method for Convex Optimization*. arXiv: 2507.16055.
- — (2025b). *The Intrinsic Riemannian Proximal Gradient Method for Nonconvex Optimization*. arXiv: 2506.09775.