LECTURE

Intrinsic First-Order Optimization: Differentials, Gradients, Constraints, and Multipliers

Basic Optimization Concepts Through the Eyes of Differential Geometry

Antonio Franchi

University of Twente and
Sapienza University of Rome

August 2026

Ver. 0.2.0 (Draft)

Guiding question

What first-order information belongs intrinsically to the cost, how does a metric turn it into a gradient, and how do equality constraints turn the same differential into a pulled-back multiplier covector?

Central message. The cost supplies an intrinsic differential. A metric turns this covector into a gradient and defines steepest descent. A constraint fiber restricts the admissible tangent directions; constrained stationarity places the same differential in the conormal space, where it becomes the pullback of a multiplier covector. Coordinates represent all these constructions and make them computable.

Use, attribution, and disclaimer

Educational use and attribution

Accuracy and responsibility

Feedback and corrections

Questions, corrections, and suggestions for improving these notes are welcome. When reporting a possible error, please indicate the relevant section, equation, proposition, or page number.

Author: Antonio Franchi

Affiliations: University of Twente and Sapienza University of Rome

Document date: August 2026

Contents

Prerequisites and scope

This lecture assumes basic familiarity with linear algebra, multivariable calculus, and the elementary language of smooth manifolds. The required concepts are summarized below to clarify what will be used without being developed from first principles.

Linear algebra

The reader is expected to be familiar with:

- finite-dimensional real vector spaces and linear subspaces;
- linear maps, kernels, images, rank, and the rank–nullity theorem;
- bases, coordinate components, and changes of basis;
- dual spaces and covectors;
- dual bases and the canonical pairing between a covector and a vector;
- dual maps, also called pullbacks in the differential-geometric setting;
- inner products, orthogonality, orthogonal complements, and positive-definite bilinear forms; and
- symmetric positive-definite matrices and their elementary properties.

In particular, the reader should be comfortable with the distinction between a vector space E and its dual space E^*. No identification between them is assumed until an inner product is introduced.

Multivariable calculus

The reader should be familiar with:

- differentiable maps between open subsets of Euclidean spaces;
- directional derivatives, Jacobian matrices, and the chain rule;
- first-order Taylor expansions;
- gradients and Hessian matrices in standard coordinates; and
- equality-constrained optimization and the classical Lagrange-multiplier equation at an introductory level.

The lecture will reinterpret several of these familiar constructions intrinsically. In particular, the usual column of partial derivatives will be reconsidered as the coordinate representation of a covector, rather than being assumed from the outset to be a vector.

Elementary differential geometry

The reader is expected to know the basic definitions of:

- smooth manifolds, charts, coordinate functions, atlases, and coordinate transitions;
- tangent spaces and tangent vectors represented by smooth curves;
- cotangent spaces and covectors;
- differentials of smooth maps and the chain rule on manifolds;
- tangent and cotangent coordinate bases;
- smooth submanifolds and inclusion maps;
- Riemannian metrics and the corresponding notions of length, angle, and orthogonality; and

- regular values and the regular level-set theorem.

Some familiarity with the tangent and cotangent bundles is useful, but only their pointwise meaning is essential. Retractions, conormal bundles, metric gradients, and the geometric interpretation of Lagrange multipliers will be introduced or reviewed when they are needed.

Conventions and level of treatment

All vector spaces are finite-dimensional and real. The manifolds $\mathcal{V}$ and $\mathcal{W}$ are assumed smooth, and all maps are assumed sufficiently smooth for the displayed derivatives to exist.

The lecture carefully separates three levels of description:

1. the intrinsic geometric objects, such as $\mathrm{d}J_v \in T_v^*\mathcal{V}$;
2. the additional structures used to represent or manipulate them, such as a Riemannian metric; and
3. the coordinate arrays used for numerical computation.

Coordinate formulas are not presented as alternatives to intrinsic geometry. They are numerical representations of the same geometric objects. Consequently, vectors, covectors, metrics, differentials, and linear maps must always be transformed according to their respective tensorial roles.

What this lecture will develop

The central constructions of the lecture are not assumed as prerequisites. The lecture will explain:
- why the differential is the primary first-order object;
- why a gradient requires a metric;
- how coordinate partial derivatives represent a covector;
- how preconditioning and Newton's direction can be interpreted through metric selection;
- why a regular equality constraint defines a fiber whose tangent space is a kernel; and
- why constrained stationarity produces a multiplier covector through the pullback of the constraint differential.

Recommended references

This lecture is intended as a compact bridge between differential geometry and first-order optimization. The geometric and algorithmic viewpoints developed here are treated in greater depth in the following references.

For optimization on smooth manifolds, including Riemannian gradients, retractions, first- and second-order algorithms, and numerical implementations, the principal companion reference is Boumal [1].

For the underlying theory of smooth manifolds, tangent and cotangent spaces, differentials, pullbacks, embedded submanifolds, regular level sets, and vector bundles, see Lee [2].

Readers interested specifically in matrix manifolds and their use in numerical optimization may also consult Absil, Mahony, and Sepulchre [3].

References

[1] N. Boumal, *An Introduction to Optimization on Smooth Manifolds*, Cambridge University Press, 2023. doi:10.1017/9781009166164.

[2] J. M. Lee, *Introduction to Smooth Manifolds*, 2nd ed., Graduate Texts in Mathematics 218, Springer, 2013. doi:10.1007/978-1-4419-9982-5.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.

Learning objectives and conceptual map

At the end of this lecture, the reader should be able to distinguish an abstract function from its coordinate formula; distinguish a differential from a gradient; characterize ascent, descent, and level directions without a metric; derive the coordinate formula for a metric gradient; interpret preconditioning and Newton's direction geometrically; formulate equality constraints as regular fibers; derive constrained stationarity through annihilators and pullbacks; explain the Jacobian transpose; distinguish conormal covectors from metric normal vectors; and separate tangent projection from a finite retraction step.

The lecture follows two enrichments of the same intrinsic covector:

$$\boxed{\mathrm{d}J_v \xrightarrow{\sharp_{g,v}} \operatorname{grad}_g J(v)} \qquad \text{and} \qquad \boxed{\mathrm{d}J_v|_{\mathrm{T}_v\mathcal{F}_w} = 0 \Longleftrightarrow \mathrm{d}J_v \in (\mathrm{T}_v\mathcal{F}_w)^\circ \Longleftrightarrow \mathrm{d}J_v \in \operatorname{Im}(\mathrm{d}f_v)^*.}$$

The first chooses a vector representative through a metric. The second restricts the directions on which the differential must vanish.

1 The function is not its formula

The function is not its formula

When we write

$$J : \mathcal{V} \to \mathbb{R},$$

the argument of J is a point of the abstract manifold $\mathcal{V}$, not a column of coordinates. A chart

$$\varphi : U \subset \mathcal{V} \to \varphi(U) \subset \mathbb{R}^n$$

produces the coordinate representative

$$J_\varphi = J \circ \varphi^{-1},$$

but this formula is only one representation of J. Another chart may make the same intrinsic function look completely different. For example, the same function on a one-dimensional manifold may be represented as

$$J_x(x) = x^2$$

in the coordinate x, and, under $y = \arctan x$, as

$$J_y(y) = \tan^2 y.$$

The formulas differ, while the assignment of a real number to each point is unchanged.
Thus neither J nor its differential has one privileged appearance. The function assigns a number to each point $v \in \mathcal{V}$; its differential at v assigns a number to each tangent vector $\xi \in \mathrm{T}_v\mathcal{V}$. Globally, $\mathrm{d}J$ is a one-form, that is, a section of $\mathrm{T}^*\mathcal{V}$.

2 The differential as the intrinsic first-order object

Let $\mathcal{V}$ be a smooth n-dimensional manifold, let a cost function $J : \mathcal{V} \to \mathbb{R}$ be smooth, and fix $v \in \mathcal{V}$.

Definition 2.1 (Differential)**.** The differential of J at v is the linear map

$$\mathrm{d}J_v : \mathrm{T}_v\mathcal{V} \to \mathbb{R}$$

defined by

$$\mathrm{d}J_v(\xi) = \frac{\mathrm{d}}{\mathrm{d}t} J(\gamma(t))\Big|_{t=0},$$

where $\gamma(0) = v$ and $\dot\gamma(0) = \xi$. Thus $\mathrm{d}J_v \in \mathrm{T}_v^*\mathcal{V}$.

Proposition 2.2 (First-order change along a curve)**.** *If $\gamma(0) = v$ and $\dot\gamma(0) = \xi$, then*

$$J(\gamma(t)) = J(v) + t\,\mathrm{d}J_v(\xi) + o(t).$$

Proof. Apply the one-variable Taylor expansion to $J \circ \gamma$ at 0, and use $(J \circ \gamma)'(0) = \mathrm{d}J_v(\xi)$. □

Geometric interpretation

A tangent vector is an admissible infinitesimal motion. A covector is a linear measuring device that accepts such a motion and returns a scalar. The value $\mathrm{d}J_v(\xi)$ is the intrinsic first-order rate of change of the cost along ξ.

A differential is not a gradient

The differential belongs to $T_v^*\mathcal{V}$, whereas a gradient belongs to $T_v\mathcal{V}$. These spaces are dual, but they are not canonically identical. Writing partial derivatives in a column does not turn the differential into a tangent vector.

3 Level, ascent, and descent directions

Assume $dJ_v \neq 0$, and define

$$\begin{aligned}\mathcal{D}_v^- &= \{\xi \in T_v\mathcal{V} : dJ_v(\xi) < 0\},\\ \mathcal{D}_v^+ &= \{\xi \in T_v\mathcal{V} : dJ_v(\xi) > 0\},\\ \mathcal{N}_v &= \ker dJ_v.\end{aligned}$$

The nonzero linear functional dJ_v has rank one, so $\mathcal{N}_v$ is a codimension-one subspace separating the two open half-spaces $\mathcal{D}_v^-$ and $\mathcal{D}_v^+$.

If $dJ_v \neq 0$, the regular-level-set theorem gives

$$T_v J^{-1}(J(v)) = \ker dJ_v.$$

Thus the kernel of dJ_v is the tangent hyperplane to the cost level set through v.

Geometric interpretation

The kernel

$$\ker dJ_v = \{\xi : dJ_v(\xi) = 0\}$$

is one linear hyperplane through the origin. More generally, each level set

$$H_a = \{\xi \in T_v\mathcal{V} : dJ_v(\xi) = a\}$$

is an affine hyperplane parallel to the kernel. These labelled hyperplanes partition $T_v\mathcal{V}$. The full labelled stack contains more information than the kernel: two covectors α and 2α have the same kernel and the same sign half-spaces, but $(2\alpha)(\xi) = 2\alpha(\xi)$. For fixed label increments, the stack of 2α appears twice as dense after a metric is chosen to visualize spacing. The labelled family is intrinsic; its drawn spacing is metric-dependent.

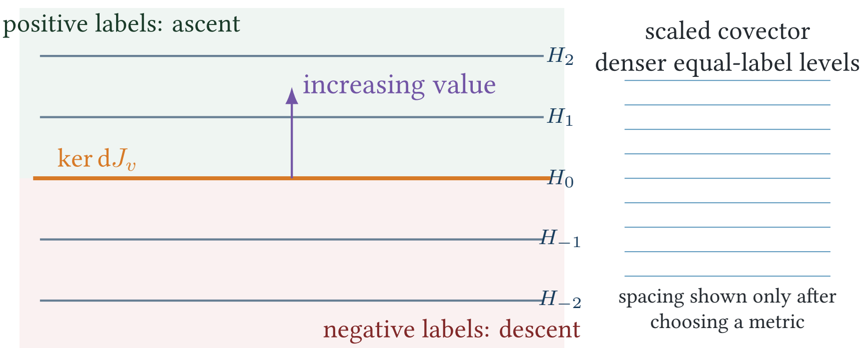


Figure 1: A nonzero covector is represented by its labelled affine level hyperplanes. The kernel is only the zero level. Scaling preserves the kernel and sign half-spaces but changes the values assigned to tangent vectors.

Definition 3.1 (Unconstrained stationary point)**.** A point $v \in \mathcal{V}$ is stationary for J if

$$\boxed{\mathrm{d}J_v = 0 \in \mathrm{T}_v^*\mathcal{V}.}$$

First-order stationarity is not optimality

The equation $\mathrm{d}J_v = 0$ identifies critical points. It does not classify them as minima, maxima, or saddles.

4 Metrics, musical maps, and gradients

Without a metric, the differential identifies all stationary points in $\mathcal{V}$ and, at every nonstationary point v, all descent directions in $T_v\mathcal{V}$, namely the set $\mathcal{D}_v^-$. It cannot, however, identify which descent direction is steepest, because unit length, angle, orthogonality, and normal direction are undefined until a metric is chosen.

Definition 4.1 (Metric, musical maps, and gradient)**.** A Riemannian metric assigns an inner product

$$g_v : \mathrm{T}_v\mathcal{V} \times \mathrm{T}_v\mathcal{V} \to \mathbb{R}.$$

It induces

$$\flat_{g,v} : \mathrm{T}_v\mathcal{V} \to \mathrm{T}_v^*\mathcal{V}, \qquad \xi \mapsto g_v(\xi, \cdot),$$

and its inverse $\sharp_{g,v}$. The metric gradient is

$$\operatorname{grad}_g J(v) = \sharp_{g,v}(\mathrm{d}J_v),$$

equivalently

$$\boxed{g_v(\operatorname{grad}_g J(v), \xi) = \mathrm{d}J_v(\xi) \qquad \forall \xi \in \mathrm{T}_v\mathcal{V}.}$$

In words, the gradient of J at a nonstationary point v is the tangent vector associated with the differential at v by the sharp map of the metric g_v. It is the unique tangent vector such that, for every other tangent vector ξ, its g_v-inner product with ξ equals the dual pairing of the differential of J with the same vector ξ. Different inner products generally associate different gradients with the same differential. Thus, the differential is unique, whereas there are infinitely many possible gradients, one for each choice of Riemannian metric.

Proposition 4.2 (Negative gradient is a descent direction)**.** *If* $\mathrm{d}J_v \neq 0$*, then*

$$\mathrm{d}J_v\big(-\operatorname{grad}_g J(v)\big) = -\left\|\operatorname{grad}_g J(v)\right\|_g^2 < 0.$$

In other words, for every Riemannian metric g*, the negative of the induced gradient is a strict descent direction:*

$$-\operatorname{grad}_g J(v) \in \mathcal{D}_v^-.$$

Proof. Substitute $\xi = -\operatorname{grad}_g J(v)$ into the defining identity. Since $\sharp_{g,v}$ is an isomorphism, a nonzero differential yields a nonzero gradient; positive definiteness gives strict negativity. ☐

Proposition 4.3 (The gradient spans the metric normal to the level hyperplane)**.** *If* $\mathrm{d}J_v \neq 0$*, then*

$$\ker \mathrm{d}J_v = \operatorname{Span}\{\operatorname{grad}_g J(v)\}^{\perp_g}. \tag{1}$$

Proof. A vector ξ belongs to the kernel exactly when $0 = \mathrm{d}J_v(\xi) = g_v(\operatorname{grad}_g J(v), \xi)$. □

Remark 4.4 (The level hyperplane precedes the metric normal). It is important to interpret Eq. (1) in the correct direction. The linear hyperplane $\ker \mathrm{d}J_v \subset T_v\mathcal{V}$, which contains the tangent directions along which the cost J remains constant to first order, is determined by the differential alone. It therefore exists independently of any metric and independently of any gradient.

The right-hand side of Eq. (1) should consequently not be understood as defining this hyperplane. Rather, once a metric g is chosen, the metric associates $\mathrm{d}J_v$ with a specific gradient vector $\operatorname{grad}_g J(v)$. The one-dimensional subspace generated by this gradient is then the g-orthogonal complement of the already existing hyperplane $\ker \mathrm{d}J_v$.

Thus, changing the metric does not change the level hyperplane. It changes the tangent-space direction that is declared perpendicular to that hyperplane and the specific gradient vector selected along that direction.

Key idea

- The hyperplane $\ker \mathrm{d}J_v \subset \mathrm{T}_v\mathcal{V}$ is determined by the cost alone.
- The metric defines $\sharp_{g,v}$ and identifies $\mathrm{d}J_v$ with $\operatorname{grad}_g J(v)$.
- If $\mathrm{d}J_v \neq 0$, the one-dimensional subspace generated by the gradient is the g-orthogonal complement of the kernel of $\mathrm{d}J_v$.
- Changing the metric does not move the kernel or alter $\mathcal{D}_v^-$ and $\mathcal{D}_v^+$.
- Changing the metric generally changes both the gradient direction and its metric-dependent norm.
- Every nonzero metric gradient lies in the same intrinsic ascent half-space $\mathcal{D}_v^+$

Proposition 4.5 (Stationarity is metric-independent). *For every Riemannian metric g,*

$$\mathrm{d}J_v = 0 \quad \Longleftrightarrow \quad \operatorname{grad}_g J(v) = 0.$$

Proof. The sharp map is a linear isomorphism, so $\sharp_{g,v}(\mathrm{d}J_v) = 0$ if and only if $\mathrm{d}J_v = 0$. □

The stationary set is invariant, while gradient fields, flows, conditioning, convergence rates, and basins of attraction can depend on the metric.

5 Variational meanings of the gradient

Proposition 5.1 (Steepest first-order ascent and descent). *If $\mathrm{d}J_v \neq 0$, then*

$$\arg \max_{\substack{\xi \in \mathrm{T}_v\mathcal{V} \\ \|\xi\|_g = 1}} \mathrm{d}J_v(\xi) = \frac{\operatorname{grad}_g J(v)}{\left\|\operatorname{grad}_g J(v)\right\|_g},$$

and the minimizer is its negative.

Proof. By Cauchy–Schwarz,

$$\mathrm{d}J_v(\xi) = g_v(\operatorname{grad}_g J(v), \xi) \leq \left\|\operatorname{grad}_g J(v)\right\|_g \|\xi\|_g .$$

For unit ξ, equality holds uniquely in the positive gradient direction; applying the same argument to $-\mathrm{d}J_v$ gives steepest descent. □

Proposition 5.2 (Gradient step as a regularized local minimizer). *For $\eta > 0$,*

$$\arg\min_{\xi \in \mathrm{T}_v\mathcal{V}} \left[\mathrm{d}J_v(\xi) + \frac{1}{2\eta} \|\xi\|_g^2 \right] = -\eta \operatorname{grad}_g J(v).$$

Proof. The directional derivative at ξ in direction ζ is

$$\mathrm{d}J_v(\zeta) + \frac{1}{\eta} g_v(\xi, \zeta) = g_v \left(\operatorname{grad}_g J(v) + \frac{1}{\eta}\xi, \zeta \right).$$

It vanishes for every ζ precisely at $\xi = -\eta \operatorname{grad}_g J(v)$. Positive definiteness gives strict convexity and uniqueness. □

Geometric interpretation

The unit-sphere problem selects only a direction in $\mathrm{T}_v\mathcal{V}$. The regularized model selects a complete tangent displacement by balancing the predicted first-order change $\mathrm{d}J_v(\xi)$ against the metric-dependent displacement cost $\frac{1}{2\eta} \|\xi\|_g^2$. Thus the metric defines both what is steepest and how local displacements are priced.

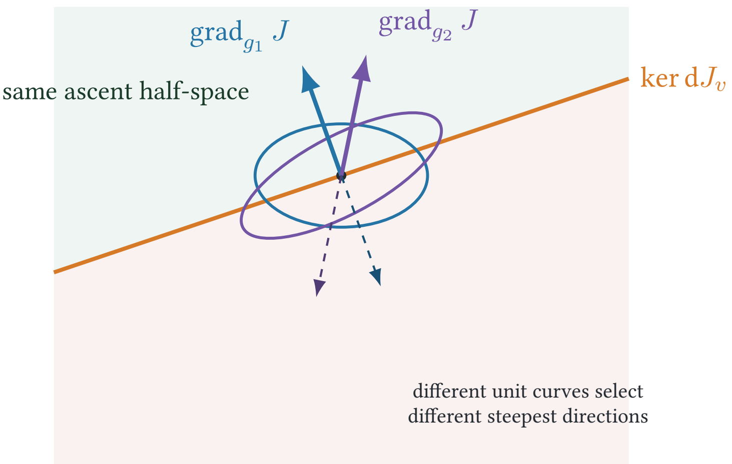


Figure 2: The level hyperplane and its ascent/descent sides are fixed by the differential. Different metrics define different unit curves and therefore different gradient directions, while every nonzero gradient remains on the ascent side.

6 Coordinates: what is actually differentiated?

To perform numerical computations, the abstract objects J, $\mathrm{d}J_v$, g_v, and $\operatorname{grad}_g J(v)$ must be represented in coordinates. Coordinates do not redefine these objects; they express them through numerical arrays suitable for storage and computation. Infinitely many coordinate systems may be chosen, producing different arrays and formulas that nevertheless represent the same intrinsic objects. Therefore, any geometric conclusion obtained in a particular coordinate system must be independent of that choice, provided that all vectors, covectors, metrics, and maps are transformed consistently.

Choose a chart

$$\varphi : U \subset \mathcal{V} \to \varphi(U) \subset \mathbb{R}^n, \qquad \varphi(q) = (x^1(q), \dots, x^n(q)).$$

It induces the tangent basis

$$\left\{ \left.\frac{\partial}{\partial x^1}\right|_v, \dots, \left.\frac{\partial}{\partial x^n}\right|_v \right\} \subset \mathrm{T}_v\mathcal{V}$$

and dual basis $\{\mathrm{d}x^1|_v, \ldots, \mathrm{d}x^n|_v\} \subset \mathrm{T}_v^*\mathcal{V}$, with

$$\mathrm{d}x^i|_v \left(\frac{\partial}{\partial x^j}\bigg|_v \right) = \delta^i_j.$$

The abstract cost J is represented by

$$J_\varphi = J \circ \varphi^{-1} : \varphi(U) \to \mathbb{R}.$$

Let $u^1, \ldots, u^n$ be the standard numerical coordinates on $\varphi(U)$. Then

$$\boxed{\mathrm{d}J_v = \sum_{i=1}^{n} \frac{\partial (J \circ \varphi^{-1})}{\partial u^i}\bigg|_{u=\varphi(v)} \mathrm{d}x^i|_v.}$$

Moreover,

$$\mathrm{d}J_v \left(\frac{\partial}{\partial x^i}\bigg|_v \right) = \frac{\partial (J \circ \varphi^{-1})}{\partial u^i}\bigg|_{u=\varphi(v)} = \frac{\partial J_\varphi}{\partial u^i}\bigg|_{u=\varphi(v)}.$$

These three expressions denote the same scalar: the i-th coordinate component of the covector $\mathrm{d}J_v$ with respect to the dual basis chosen above. This scalar is conventionally abbreviated as

$$\frac{\partial J}{\partial x^i}(v).$$

What is actually being differentiated?

The notation $\partial J/\partial x^i$ does not mean that the abstract cost J intrinsically comes with numerical variables. Ordinary differentiation rules are applied to $J \circ \varphi^{-1}$. The resulting numbers are the components of the intrinsic covector $\mathrm{d}J_v$ in the basis $\{\mathrm{d}x^i|_v\}$. Both components and basis change under a new chart, while the covector does not.

Store the components as

$$d_v := \begin{bmatrix} \partial J/\partial x^1(v) \\ \vdots \\ \partial J/\partial x^n(v) \end{bmatrix}.$$

This is a column of covector components. Column orientation is a storage convention; it does not change tensor type.

7 Coordinate gradients and tensorial covariance

Let

$$G_v = [g_{ij}(v)], \qquad g_{ij}(v) = g_v \left(\frac{\partial}{\partial x^i}\bigg|_v, \frac{\partial}{\partial x^j}\bigg|_v \right).$$

Write

$$\mathrm{grad}_g\, J(v) = \sum_{i=1}^{n} \zeta^i \frac{\partial}{\partial x^i}\bigg|_v, \qquad [\mathrm{grad}_g\, J]_v = \begin{bmatrix} \zeta^1 \\ \vdots \\ \zeta^n \end{bmatrix}.$$

For an arbitrary tangent vector ξ

$$\xi = \sum_{j=1}^{n} \xi^j \left.\frac{\partial}{\partial x^j}\right|_v, \qquad [\xi]_v = \begin{bmatrix} \xi^1 \\ \vdots \\ \xi^n \end{bmatrix},$$

the gradient defining identity

$$g_v\big(\mathrm{grad}_g\, J(v), \xi\big) = \mathrm{d}J_v(\xi)$$

becomes

$$[\mathrm{grad}_g\, J]_v^\top G_v [\xi]_v = d_v^\top [\xi]_v.$$

Since this identity holds for every component vector $[\xi]_v$, and since G_v is symmetric, it follows that

$$G_v [\mathrm{grad}_g\, J]_v = d_v.$$

Therefore,

$$\boxed{[\mathrm{grad}_g\, J]_v = G_v^{-1} d_v.}$$

Coordinate realization

The arrays d_v and $[\mathrm{grad}_g\, J]_v$ represent geometrically different objects:

- d_v contains the components of the covector $\mathrm{d}J_v$ in the coordinate cotangent basis $\{\mathrm{d}x^i|_v\}$;
- $[\mathrm{grad}_g\, J]_v$ contains the components of the tangent vector $\mathrm{grad}_g\, J(v)$ in the coordinate tangent basis $\{\partial/\partial x^i|_v\}$;
- G_v^{-1} represents the sharp map that converts the covector components into the corresponding vector components:

$$[\mathrm{grad}_g\, J]_v = G_v^{-1} d_v.$$

The components of a covector are conventionally called *covariant components*, whereas the components of a tangent vector are called *contravariant components*. These names refer to their different transformation laws under a change of coordinates.

Let $y = \psi(x)$ be another coordinate system and set

$$K = \frac{\partial y}{\partial x}.$$

Then

$$[\xi]_y = K[\xi]_x, \qquad [\mathrm{d}J]_y = K^{-\top}[\mathrm{d}J]_x,$$

so the pairing is invariant:

$$[\mathrm{d}J]_y^\top [\xi]_y = [\mathrm{d}J]_x^\top K^{-1} K [\xi]_x = [\mathrm{d}J]_x^\top [\xi]_x.$$

The metric and inverse metric transform as

$$G_y = K^{-\top} G_x K^{-1}, \qquad G_y^{-1} = K G_x^{-1} K^\top,$$

and therefore

$$[\mathrm{grad}_g\, J]_y = G_y^{-1}[\mathrm{d}J]_y = K[\mathrm{grad}_g\, J]_x.$$

The gradient consequently transforms as a vector.

Example 7.1 (Euclidean geometry in polar coordinates). For $J(x, y) = \frac{1}{2}(x^2 + y^2)$, Cartesian coordinates give $G_{xy} = I$ and $[\mathrm{d}J]_{xy} = (x, y)^\top$. In polar coordinates,

$$J(r,\theta) = \tfrac{1}{2}r^2, \qquad [\mathrm{d}J]_{r\theta} = \begin{bmatrix} r \\ 0 \end{bmatrix}, \qquad G_{r\theta} = \begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix}.$$

Thus $[\operatorname{grad} J]_{r\theta} = (r, 0)^\top$, representing $r\,\partial/\partial r$, the same radial vector as in Cartesian coordinates.

The identity matrix is not coordinate-invariant

Setting the metric matrix to I again after a nonlinear coordinate change generally chooses a different metric. In polar coordinates the Euclidean metric is $\operatorname{diag}(1, r^2)$, not I.

8 Changing the metric and preconditioning

We have seen that the differential $\mathrm{d}J_v$ is fixed by the cost, whereas the gradient depends on the metric used to convert this covector into a tangent vector. We now make explicit how the gradient changes when the metric changes, first intrinsically and then in coordinates.

8.1 Intrinsic comparison of two metrics

Let g and h be two Riemannian metrics on $\mathcal{V}$. At each $v \in \mathcal{V}$, define the linear map

$$A_v : T_v\mathcal{V} \to T_v\mathcal{V}$$

by

$$A_v = \sharp_{g,v} \circ \flat_{h,v}.$$

Equivalently, A_v is the unique map satisfying

$$h_v(\xi, \zeta) = g_v\big(A_v(\xi), \zeta\big) \qquad \forall \xi, \zeta \in T_v\mathcal{V}.$$

The map A_v expresses the metric h_v in terms of the metric g_v. Such a map exists and is unique for every pair of inner products because the flat map of h_v sends a tangent vector to a covector, while the sharp map of g_v sends that covector back to a tangent vector:

$$T_v\mathcal{V} \xrightarrow{\flat_{h,v}} T_v^*\mathcal{V} \xrightarrow{\sharp_{g,v}} T_v\mathcal{V}.$$

The metric-change operator A_v is self-adjoint and positive with respect to g_v:

$$g_v\big(A_v(\xi), \zeta\big) = g_v\big(\xi, A_v(\zeta)\big),$$

and

$$g_v\big(A_v(\xi), \xi\big) > 0 \qquad \text{for every } \xi \neq 0.$$

These properties follow directly from the symmetry and positive definiteness of h_v.

Proposition 8.1 (How the gradient changes with the metric). *The gradients of J associated with g and h satisfy*

$$\boxed{\operatorname{grad}_h J(v) = A_v^{-1}\big(\operatorname{grad}_g J(v)\big).}$$

Proof. For every $\zeta \in T_v\mathcal{V}$, $g_v\big(\operatorname{grad}_g J(v), \zeta\big) = \mathrm{d}J_v(\zeta) = h_v\big(\operatorname{grad}_h J(v), \zeta\big) = g_v\big(A_v(\operatorname{grad}_h J(v)), \zeta\big)$. Since this identity holds for every ζ, the nondegeneracy of g_v gives $\operatorname{grad}_g J(v) = A_v\big(\operatorname{grad}_h J(v)\big)$. Because A_v is invertible, $\operatorname{grad}_h J(v) = A_v^{-1}\big(\operatorname{grad}_g J(v)\big)$. □

Geometric interpretation

Changing the metric does not modify the differential $\mathrm{d}J_v$. It modifies the sharp map used to convert that fixed covector into a tangent vector. The metric-change operator A_v describes the relation between the two resulting gradients.

8.2 Coordinate viewpoint: preconditioning the differential

We now temporarily reason entirely in coordinates. Let d_v be the component column representing the covector $\mathrm{d}J_v$. Choose a symmetric positive-definite matrix

$$P_v = P_v^\top \succ 0$$

and define the component column

$$[\xi]_v = -P_v d_v.$$

At this point, the multiplication must be given a geometric interpretation. Since d_v contains covector components, we interpret P_v as the coordinate matrix of a map from covectors to vectors. Equivalently, P_v represents an inverse metric g_v^{-1} with coordinate representation G_v^{-1}:

$$P_v = G_v^{-1}, \qquad G_v = P_v^{-1}.$$

The resulting column therefore represents the tangent vector

$$\xi = -\operatorname{grad}_g J(v).$$

The intrinsic descent test is obtained by evaluating the original differential on this vector:

$$\mathrm{d}J_v(\xi) = d_v^\top [\xi]_v = -d_v^\top P_v d_v.$$

Because P_v is positive definite,

$$\mathrm{d}J_v(\xi) < 0 \qquad \text{whenever } d_v \neq 0.$$

Thus the vector represented by $-P_v d_v$ always belongs to the intrinsic descent half-space:

$$\xi \in \mathcal{D}_v^-.$$

Definition 8.2 (Preconditioning)**.** In first-order optimization, *preconditioning* consists of multiplying the component column of the differential by a chosen positive-definite matrix:

$$[\xi]_v = -P_v d_v.$$

When P_v is interpreted as an inverse metric matrix, the preconditioned direction is precisely the negative gradient associated with the metric represented by

$$G_v = P_v^{-1}.$$

Key idea

Preconditioning does not modify the intrinsic differential $\mathrm{d}J_v$. It changes the covector-to-vector identification used to produce a descent direction from that differential. In geometric terms, choosing a preconditioner is choosing a metric.

Why inverse-metric maps cannot be stacked blindly

Suppose

$$[\xi_1]_v = -P_1 d_v,$$

where P_1 is interpreted as an inverse metric. The input d_v contains covector components, while the output $[\xi_1]_v$ contains vector components.

Applying a second inverse metric P_2 directly to $[\xi_1]_v$ is therefore not type-correct: an inverse metric maps covectors to vectors, not vectors to vectors. If P_2 is instead reinterpreted as a tangent-space endomorphism, the multiplication is type-correct and generates the coordinates of another tangent vector ξ_2, namely $[\xi_2]_v = -P_2 P_1 d_v$, but it is no longer another metric identification and need not preserve descent. Indeed,

$$\mathrm{d}J_v(\xi_2) = -d_v^\top P_2 P_1 d_v$$

need not be negative because the product of two symmetric positive-definite matrices is not generally symmetric positive definite.

A new metric should act directly on the original differential:

$$[\xi_{\text{new}}]_v = -P_{\text{new}} d_v,$$

not on a vector produced by a previous metric identification.

8.3 Newton's method as Hessian-based preconditioning

The preceding discussion shows that any symmetric positive-definite matrix P_v may be interpreted as an inverse metric and used to convert the covector components d_v into the components of a descent vector:

$$[\xi]_v = -P_v d_v.$$

The resulting direction depends on how P_v, or equivalently the metric matrix $G_v = P_v^{-1}$, is chosen.

Newton's method makes a particular, cost-dependent choice. In Euclidean coordinates, let

$$H_v := \nabla^2 J(v)$$

denote the Hessian matrix of the coordinate representation of the cost. When

$$H_v = H_v^\top \succ 0,$$

Newton chooses

$$P_v = H_v^{-1}.$$

The Newton direction is therefore

$$\boxed{[\xi_{\mathrm{N}}]_v = -H_v^{-1} d_v.}$$

In Euclidean coordinates, where the component column d_v is numerically identical to the usual Euclidean-gradient column, this is the familiar Newton equation

$$H_v [\xi_{\mathrm{N}}]_v = -d_v,$$

or, in conventional notation,

$$\nabla^2 J(v)\, \xi_{\mathrm{N}} = -\nabla J(v).$$

The positive-definite Hessian matrix can also be interpreted as the matrix of a local inner product:

$$g_v^{\mathrm{N}}(\xi, \zeta) = [\xi]_v^\top H_v [\zeta]_v.$$

Its inverse metric is represented by H_v^{-1}. Consequently,

$$[\operatorname{grad}_{g^{\mathrm{N}}} J]_v = H_v^{-1} d_v,$$

and the Newton direction is precisely the negative gradient associated with this Hessian-based metric:

$$\boxed{\xi_{\mathrm{N}} = -\operatorname{grad}_{g^{\mathrm{N}}} J(v).}$$

The geometric interpretation of Newton's method

Newton's method uses the Hessian to select the preconditioner

$$P_v = H_v^{-1}.$$

Equivalently, it uses the Hessian to select a local metric and then converts the differential into a gradient through the sharp map of that metric.
Newton's method is a second-order method because constructing this metric requires second derivatives of the cost. Its direction is nevertheless a gradient direction in the Hessian-based geometry:

Newton uses second-order information to choose the geometry in which it takes a gradient step.

This viewpoint also qualifies the usual comparison between "gradient descent" and "Newton's method." There is no metric-free or uniquely natural gradient descent: every gradient is obtained by using a chosen metric to convert the differential into a tangent vector. What is conventionally called gradient descent usually means gradient descent under the metric represented by the identity matrix in the selected coordinates. When the Hessian is positive definite, Newton's method instead uses the Hessian as a local metric. In this sense, both methods are metric-gradient methods; their essential difference is the geometry used to select the descent direction and the information required to construct that geometry.

8.3.1 Why the Hessian metric can improve convergence

In accordance with Proposition 5.2, the gradient step associated with a generic metric represented by G_v minimizes the regularized first-order model

$$[\xi]_v \longmapsto d_v^\top [\xi]_v + \frac{1}{2\eta} [\xi]_v^\top G_v [\xi]_v.$$

If we choose

$$\eta = 1, \qquad G_v = H_v,$$

this model becomes

$$[\xi]_v \longmapsto d_v^\top [\xi]_v + \frac{1}{2} [\xi]_v^\top H_v [\xi]_v.$$

Apart from the constant term $J(v)$, this is exactly the local second-order quadratic model of the cost. Its unique minimizer is

$$[\xi_{\mathrm{N}}]_v = -H_v^{-1} d_v.$$

This comparison reveals the essential difference between choosing the metric represented by the identity matrix in the selected coordinates and choosing the Hessian metric. With the identity-matrix choice, the regularization term is a multiple of

$$[\xi]_v^\top [\xi]_v.$$

All coordinate directions are therefore treated as equally expensive, regardless of how rapidly the cost bends along them.

With the Hessian-metric choice, the displacement cost becomes

$$[\xi]_v^\top H_v [\xi]_v.$$

Coordinate directions of large positive curvature are penalized more strongly, whereas coordinate directions of small positive curvature are penalized less. The resulting gradient direction compensates for the local anisotropy of the cost. This curvature-dependent rescaling can reduce the slow zigzagging behavior produced by the identity-metric gradient direction near strongly elongated level sets.

What is actually being compared?

The comparison is not between a uniquely defined "gradient descent" method and a fundamentally different Newton method. Every gradient direction depends on a chosen metric.
The usual coordinate formula

$$-d_v$$

implicitly uses the metric represented by the identity matrix in the selected coordinates. The Newton direction

$$-H_v^{-1} d_v$$

uses instead the metric represented by the positive-definite Hessian H_v. Both directions are obtained by converting the same differential into a tangent vector; they differ in the geometry used for that conversion. Newton's method remains second order because the Hessian metric is constructed from second derivatives.

Example 8.3 (Exact positive-definite quadratic)**.** Consider the coordinate domain $\mathbb{R}^n$ and the quadratic cost

$$J(z) = \frac{1}{2} z^\top Q z, \qquad Q = Q^\top \succ 0.$$

The differential is represented by

$$d_z = Qz,$$

and the Hessian is the constant matrix

$$H_z = Q.$$

If the selected metric is represented by the identity matrix, the negative gradient direction has components

$$-[\operatorname{grad}_I J]_z = -d_z = -Qz.$$

If Q is ill-conditioned, the level sets of J are strongly elongated ellipsoids. Gradient iterations based on this metric may oscillate between the steep sides of the ellipsoids and require a small step size for stability.

If the selected metric is instead represented by the Hessian,

$$G = Q,$$

then

$$[\operatorname{grad}_Q J]_z = Q^{-1} d_z = Q^{-1} Q z = z.$$

The corresponding negative gradient is

$$-[\operatorname{grad}_Q J]_z = -z,$$

which is precisely the Newton direction. Since this example is posed on the vector space $\mathbb{R}^n$, a unit coordinate step gives

$$z_+ = z - z = 0.$$

Thus the minimizer is reached in one iteration.

Geometrically, the Hessian metric compensates exactly for the unequal curvatures of this quadratic. The strongly elongated level sets seen under the identity-metric representation become isotropic with respect to the Hessian metric used to select the step.

When the Hessian cannot be used as a metric

The Hessian-based interpretation requires

$$H_v \succ 0.$$

If H_v is indefinite, it does not define an inner product and therefore cannot represent a Riemannian metric. In that case,

$$-d_v^\top H_v^{-1} d_v$$

need not be negative, so the unmodified Newton direction need not be a descent direction. Damped Newton methods, positive-definite Hessian modifications, trust-region methods, and quasi-Newton methods address this difficulty by controlling or replacing the local curvature model. A further geometric qualification is necessary. The matrix of ordinary second coordinate derivatives is not, by itself, invariant under a nonlinear change of coordinates. On a manifold, an intrinsic Hessian requires additional geometric structure, typically a connection. The coordinate discussion above therefore applies directly to a chosen local coordinate model. An intrinsic manifold formulation uses a Riemannian or covariant Hessian, and the Hessian-metric interpretation applies where the resulting symmetric bilinear form is positive definite.

9 Equality constraints as regular fibers

Let $\mathcal{V}$ and $\mathcal{W}$ be two smooth manifolds called the total space and the base space (or task space), respectively. Let

$$f : \mathcal{V} \to \mathcal{W}$$

be smooth, let $w \in \mathcal{W}$ be a regular value, and define the fiber

$$\mathcal{F}_w = f^{-1}(w).$$

Theorem 9.1 (Regular fiber). *The set $\mathcal{F}_w$ is an embedded submanifold and, for every $v \in \mathcal{F}_w$,*

$$\boxed{\mathrm{T}_v\mathcal{F}_w = \ker \mathrm{d}f_v.}$$

Proof sketch. A curve in the fiber satisfies $f(\gamma(t)) = w$, so differentiation gives $\mathrm{d}f_v(\dot\gamma(0)) = 0$, proving inclusion in the kernel. Regularity and the regular-level-set theorem give equality. □

Geometric interpretation

The nonlinear equality $f(v) = w$ becomes the linear first-order condition $\mathrm{d}f_v(\xi) = 0$. The kernel of $\mathrm{d}J_v$ contains directions that do not change the cost to first order; the kernel of $\mathrm{d}f_v$ contains directions that do not change the task value to first order.

10 Constrained stationarity through the inclusion map

Let

$$\iota_w : \mathcal{F}_w \hookrightarrow \mathcal{V}$$

be the inclusion. Then

$$J|_{\mathcal{F}_w} = J \circ \iota_w$$

and the chain rule gives

$$\mathrm{d}(J|_{\mathcal{F}_w})_v = (\mathrm{d}\iota_{w,v})^*(\mathrm{d}J_v).$$

Hence

$$\boxed{\begin{aligned} v \text{ stationary for } J|_{\mathcal{F}_w} &\iff \mathrm{d}(J|_{\mathcal{F}_w})_v = 0 \\ &\iff (\mathrm{d}\iota_{w,v})^*(\mathrm{d}J_v) = 0 \\ &\iff \mathrm{d}J_v(\xi) = 0 \quad \forall \xi \in \mathrm{T}_v\mathcal{F}_w \\ &\iff \mathrm{d}J_v \in (\mathrm{T}_v\mathcal{F}_w)^\circ. \end{aligned}}$$

Thus unconstrained stationarity requires $\mathrm{d}J_v$ to vanish on all of $\mathrm{T}_v\mathcal{V}$, whereas constrained stationarity requires it to vanish only on feasible tangent directions.

11 Annihilators, dual maps, and multipliers

For a finite-dimensional linear map $L : E \to F$, define

$$L^* : F^* \to E^*, \qquad L^*(\mu) = \mu \circ L.$$

Lemma 11.1 (Annihilator identity).

$$\boxed{\operatorname{Im} L^* = (\ker L)^\circ.}$$

Proof. If $\alpha = L^*\mu$ and $e \in \ker L$, then $\alpha(e) = \mu(Le) = 0$, so $\operatorname{Im} L^* \subseteq (\ker L)^\circ$. In finite dimensions, $\dim \operatorname{Im} L^* = \operatorname{rank} L^* = \operatorname{rank} L$, while $\dim(\ker L)^\circ = \dim E - \dim \ker L = \operatorname{rank} L$. The inclusion between equal-dimensional subspaces is equality. □

Applying the lemma to $L = \mathrm{d}f_v$ gives

$$(\mathrm{T}_v\mathcal{F}_w)^\circ = (\ker \mathrm{d}f_v)^\circ = \operatorname{Im}(\mathrm{d}f_v)^*.$$

Theorem 11.2 (Intrinsic Lagrange multiplier theorem). *A point $v \in \mathcal{F}_w$ is stationary for $J|_{\mathcal{F}_w}$ if and only if there exists a unique covector*

$$\mu \in \mathrm{T}_w^*\mathcal{W}$$

such that

$$\boxed{\mathrm{d}J_v = (\mathrm{d}f_v)^*\mu.}$$

Proof. Constrained stationarity is equivalent to $\mathrm{d}J_v \in (\mathrm{T}_v\mathcal{F}_w)^\circ$, and the annihilator identity identifies this space with $\operatorname{Im}(\mathrm{d}f_v)^*$. Regularity makes $\mathrm{d}f_v$ surjective, hence $(\mathrm{d}f_v)^*$ injective, so μ is unique. □

Key idea

The multiplier equation is not initially a linear dependence among gradient vectors. The multiplier μ is a covector on the task-value space. Its pullback is the cost covector in the total-space cotangent space.

11.1 Conormal bundle in the cotangent bundle

The conormal bundle of the embedded fiber $\mathcal{F}_w \subset \mathcal{V}$ is a smooth vector subbundle of the ambient cotangent bundle restricted to $\mathcal{F}_w$:

$$N^*\mathcal{F}_w \subset T^*\mathcal{V}|_{\mathcal{F}_w} \subset T^*\mathcal{V}.$$

It is defined by

$$N^*\mathcal{F}_w = \bigsqcup_{v \in \mathcal{F}_w} N_v^*\mathcal{F}_w, \qquad N_v^*\mathcal{F}_w := (T_v\mathcal{F}_w)^\circ.$$

Equivalently,

$$N^*\mathcal{F}_w = \{(v,\alpha) \in T^*\mathcal{V} : v \in \mathcal{F}_w,\ \alpha(\xi) = 0 \text{ for every } \xi \in T_v\mathcal{F}_w\}.$$

For every $v \in \mathcal{F}_w$, the differential $\mathrm{d}J_v \in T_v^*\mathcal{V}$ determines a point

$$(v, \mathrm{d}J_v) \in T^*\mathcal{V}.$$

This point belongs to the conormal bundle precisely when $\mathrm{d}J_v$ vanishes on every feasible tangent direction:

$$\boxed{(v, \mathrm{d}J_v) \in N^*\mathcal{F}_w \quad \Longleftrightarrow \quad \mathrm{d}J_v(\xi) = 0 \quad \forall \xi \in T_v\mathcal{F}_w.}$$

Consequently, the constrained stationary points are exactly those points $v \in \mathcal{F}_w$ at which the cost differential belongs to the corresponding conormal fiber $N_v^*\mathcal{F}_w$. Depending on the cost and the constraint fiber, there may be no such points, a single point, or several points.

The tangent bundle of $\mathcal{F}_w$ collects the feasible infinitesimal motions, whereas the conormal bundle collects the covectors that vanish on those motions. Constrained stationarity occurs when the differential of the cost is one of these conormal covectors.

12 Lagrangian multipliers and bridge with the intrinsic formulation

Suppose first that $\mathcal{W}$ is a vector space, and let $\lambda \in \mathcal{W}^*$ be a task-space covector. The conventional plus-sign Lagrangian is

$$\boxed{\mathcal{L}(v,\lambda;w) = J(v) + \langle \lambda, f(v) - w \rangle.}$$

On a general manifold $\mathcal{W}$, the subtraction $f(v) - w$ is not intrinsically defined; the same expression must be understood through a local chart or a vector-valued constraint residual. Differentiating with respect to v yields

$$\mathrm{d}_v\mathcal{L} = \mathrm{d}J_v + (\mathrm{d}f_v)^*\lambda.$$

Thus stationarity gives

$$\boxed{\mathrm{d}J_v + (\mathrm{d}f_v)^*\lambda = 0.}$$

Define once and for all

$$\boxed{\mu := -\lambda.}$$

Then the conventional Lagrangian equation is exactly the geometric pullback equation $\mathrm{d}J_v = (\mathrm{d}f_v)^*\mu$.

13 Coordinates of the task differential and multiplier

Choose charts

$$\varphi : U \subset \mathcal{V} \to \mathbb{R}^n, \qquad \psi : Z \subset \mathcal{W} \to \mathbb{R}^m,$$

and define the coordinate representative

$$f_{\psi\varphi} = \psi \circ f \circ \varphi^{-1}.$$

The matrix representing $\mathrm{d}f_v$ is

$$F_v = D(\psi \circ f \circ \varphi^{-1})(\varphi(v)).$$

Hence

$$[\mathrm{d}f_v(\xi)]_\psi = F_v[\xi]_\varphi.$$

For $\mu \in \mathrm{T}^*_w\mathcal{W}$,

$$\begin{aligned} ((\mathrm{d}f_v)^*\mu)(\xi) &= \mu(\mathrm{d}f_v(\xi)) \\ &= [\mu]_\psi^\top F_v[\xi]_\varphi \\ &= (F_v^\top[\mu]_\psi)^\top[\xi]_\varphi. \end{aligned}$$

Therefore

$$\boxed{[(\mathrm{d}f_v)^*\mu]_\varphi = F_v^\top[\mu]_\psi.}$$

The intrinsic equations become

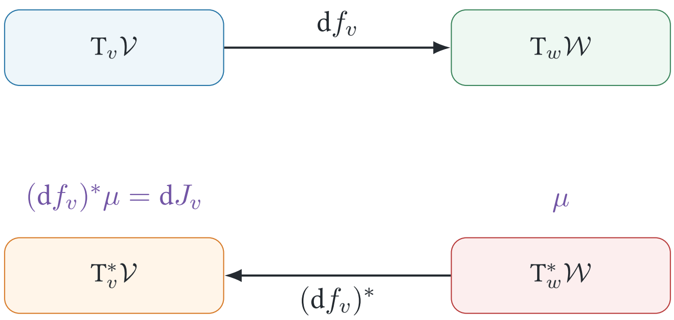
$$\boxed{d_v = F_v^\top[\mu]_\psi} \qquad \text{and} \qquad \boxed{d_v + F_v^\top[\lambda]_\psi = 0.}$$

Under Euclidean musical identifications,

$$\boxed{\nabla J(v) + F_v^\top\lambda = 0.}$$

Coordinate realization

The transpose is forced by duality. The differential pushes tangent vectors forward through F_v; the dual map pulls covectors backward through $F_v^\top$. It is not an accidental algebraic operation.



14 Conormal covectors and metric normals

Define the metric normal space

$$N^g_v\mathcal{F}_w = (\mathrm{T}_v\mathcal{F}_w)^{\perp_g}.$$

Proposition 14.1 (Sharp maps conormals to metric normals)**.**

$$\boxed{\sharp_{g,v}\big((\mathrm{T}_v\mathcal{F}_w)^\circ\big) = N_v^g\mathcal{F}_w.}$$

Proof. If α annihilates $\mathrm{T}_v\mathcal{F}_w$ and $u = \sharp_{g,v}\alpha$, then for every $\xi \in \mathrm{T}_v\mathcal{F}_w$,

$$g_v(u, \xi) = \alpha(\xi) = 0,$$

so u is normal. Conversely, the flat image of every metric-normal vector annihilates the tangent space. □

Thus

$$\operatorname{grad}_g J(v) = \sharp_{g,v}((\mathrm{d}f_v)^*\mu) \in N_v^g\mathcal{F}_w,$$

or, with $\lambda = -\mu$,

$$\operatorname{grad}_g J(v) + \sharp_{g,v}((\mathrm{d}f_v)^*\lambda) = 0.$$

Normality requires a metric

The phrase "the gradient is normal to the constraint" already presupposes a metric. The primary invariant statement is that $\mathrm{d}J_v$ annihilates $\mathrm{T}_v\mathcal{F}_w$.

15 Running example: optimization on the sphere and the Rayleigh quotient

Let $\mathcal{V} = \mathbb{R}^n$ be equipped with its standard inner product, let $\mathcal{W} = \mathbb{R}$, and define

$$f(v) = \frac{1}{2}\langle v, v\rangle, \qquad w = \frac{1}{2}.$$

The corresponding fiber is the unit sphere,

$$\mathcal{F}_{1/2} = f^{-1}\left(\frac{1}{2}\right) = \mathbb{S}^{n-1}.$$

For every $\xi \in T_v\mathbb{R}^n \simeq \mathbb{R}^n$, one has $\mathrm{d}f_v(\xi) = \langle v, \xi\rangle$. Hence

$$T_v\mathbb{S}^{n-1} = \ker \mathrm{d}f_v = \{\xi \in \mathbb{R}^n : \langle v, \xi\rangle = 0\},$$

so the feasible tangent space is the hyperplane orthogonal to v.

A point $v \in \mathbb{S}^{n-1}$ is stationary for $J|_{\mathbb{S}^{n-1}}$ if and only if $\mathrm{d}J_v(\xi) = 0$ for every $\xi \in T_v\mathbb{S}^{n-1}$. Equivalently, there exists a geometric multiplier $\mu \in T^*_{1/2}\mathbb{R} \simeq \mathbb{R}$ such that

$$\boxed{\mathrm{d}J_v = \mu\, \mathrm{d}f_v.}$$

With the conventional plus-sign Lagrangian

$$\mathcal{L}(v, \lambda) = J(v) + \lambda\left(\frac{1}{2}\langle v, v\rangle - \frac{1}{2}\right),$$

the same condition is

$$\mathrm{d}J_v + \lambda\, \mathrm{d}f_v = 0, \qquad \lambda = -\mu.$$

Using the standard inner product to identify covectors with vectors, this becomes

$$\operatorname{grad} J(v) + \lambda v = 0.$$

15.1 A self-adjoint quadratic cost

Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be self-adjoint with respect to the standard inner product. Thus $\langle A\xi, \zeta\rangle = \langle \xi, A\zeta\rangle$ for all $\xi, \zeta \in \mathbb{R}^n$. In an orthonormal basis, A is represented by a symmetric matrix, also denoted by A.

Consider

$$J(v) = \frac{1}{2}\langle v, Av\rangle.$$

Its differential is $\mathrm{d}J_v(\xi) = \langle Av, \xi\rangle$. The intrinsic multiplier equation therefore gives

$$\langle Av, \xi\rangle = \mu\langle v, \xi\rangle \qquad \forall \xi \in \mathbb{R}^n.$$

By nondegeneracy of the inner product,

$$\boxed{Av = \mu v.}$$

Consequently, $v \in \mathbb{S}^{n-1}$ is stationary for $J|_{\mathbb{S}^{n-1}}$ if and only if v is a unit eigenvector of A. The geometric multiplier μ is the corresponding eigenvalue, while the plus-sign Lagrangian multiplier is $\lambda = -\mu$.

15.2 Rayleigh quotient and stationary set

The Rayleigh quotient of A is

$$\mathcal{R}_A(v) = \frac{\langle v, Av\rangle}{\langle v, v\rangle}, \qquad v \neq 0.$$

It is invariant under nonzero rescaling: $\mathcal{R}_A(cv) = \mathcal{R}_A(v)$ for every $c \neq 0$. On the unit sphere,

$$\mathcal{R}_A(v) = \langle v, Av\rangle = 2J(v).$$

Therefore, the restrictions of J and $\mathcal{R}_A$ to the sphere have the same stationary points. If $Av = \mu v$ and $\langle v, v\rangle = 1$, then

$$\mathcal{R}_A(v) = \mu, \qquad J(v) = \frac{\mu}{2}.$$

Thus the geometric multiplier is the stationary value of the Rayleigh quotient.

Let $E_\rho = \ker(A - \rho I)$ be the eigenspace associated with ρ. The complete constrained stationary set is

$$\boxed{\operatorname{Crit}\left(J|_{\mathbb{S}^{n-1}}\right) = \bigcup_{\rho \in \sigma(A)} \left(E_\rho \cap \mathbb{S}^{n-1}\right).}$$

Hence every unit vector in an eigenspace is stationary. Simple eigenvalues produce antipodal pairs $\pm e_i$; an eigenvalue of higher multiplicity produces an entire unit sphere within the corresponding eigenspace.

Moreover,

$$\min_{v \neq 0} \mathcal{R}_A(v) = \lambda_{\text{min}}(A), \qquad \max_{v \neq 0} \mathcal{R}_A(v) = \lambda_{\text{max}}(A),$$

attained on the unit vectors in the corresponding extremal eigenspaces. The minimum and maximum values of J on the unit sphere are therefore $\lambda_{\text{min}}(A)/2$ and $\lambda_{\text{max}}(A)/2$, respectively.

Geometric interpretation

At $v \in \mathbb{S}^{n-1}$, the feasible tangent vectors are precisely those orthogonal to v. Under the chosen inner product, the cost differential is represented by Av. Constrained stationarity requires

$$\langle Av, \xi \rangle = 0 \qquad \forall \xi \in T_v\mathbb{S}^{n-1}.$$

Thus Av must be orthogonal to the entire tangent hyperplane and must consequently belong to its one-dimensional normal space:

$$Av \in \mathrm{Span}\{v\}.$$

This is exactly the eigenvector condition. The stationary points are therefore the points at which the gradient of the quadratic cost is normal to the sphere.

Remark 15.1 (Intrinsic meaning of A and its eigenvalues). The primary object is the self-adjoint endomorphism $A : \mathbb{R}^n \to \mathbb{R}^n$, not its matrix representation. Under a change of basis, the representing matrix changes by similarity, whereas the eigenvalues and eigenspaces of the operator remain unchanged. In an orthonormal basis, self-adjointness is represented by the familiar matrix condition $A^\top = A$.

If one starts instead from a symmetric bilinear form, an inner product is needed to associate that form with a self-adjoint endomorphism before its eigenvalues can be interpreted intrinsically.

16 The multiplier as the differential of the optimal value

We now show that the geometric multiplier measures how the optimal cost changes when the prescribed task value w is varied.

Let $U \subset \mathcal{W}$ be a neighborhood of w. Assume that, for every $w' \in U$, the constrained problem

$$\min_{v \in \mathcal{F}_{w'}} J(v), \qquad \mathcal{F}_{w'} = f^{-1}(w'),$$

has a selected minimizer

$$\widehat{v}(w') \in \mathcal{F}_{w'},$$

and that the resulting map

$$\widehat{v} : U \to \mathcal{V}, \qquad w' \mapsto \widehat{v}(w'),$$

is differentiable.

Define the optimal-value function along this differentiable selection of minimizers by

$$\mathcal{J}^{\mathrm{opt}}(w') := J\big(\widehat{v}(w')\big) = \min_{v \in \mathcal{F}_{w'}} J(v).$$

The feasibility of the selected minimizer gives

$$f\big(\widehat{v}(w')\big) = w'.$$

For a task variation $\delta w \in T_w\mathcal{W}$, differentiating this identity at w and applying the chain rule yields

$$\mathrm{d}f_{\widehat{v}(w)}\left(\mathrm{d}\widehat{v}_w[\delta w]\right) = \delta w.$$

Let

$$\mu \in T_w^*\mathcal{W}$$

be the geometric multiplier at $\widehat{v}(w)$, so that

$$\mathrm{d}J_{\widehat{v}(w)} = \left(\mathrm{d}f_{\widehat{v}(w)}\right)^{*}\mu.$$

Applying the chain rule to the optimal-value function gives

$$\begin{aligned}\mathrm{d}\mathcal{J}_w^{\mathrm{opt}}(\delta w) &= \mathrm{d}J_{\widehat{v}(w)}\left(\mathrm{d}\widehat{v}_w[\delta w]\right) \\ &= \mu\left(\mathrm{d}f_{\widehat{v}(w)}\left(\mathrm{d}\widehat{v}_w[\delta w]\right)\right) \\ &= \mu(\delta w).\end{aligned}$$

Since this equality holds for every $\delta w \in T_w\mathcal{W}$, we conclude that

$$\boxed{\mathrm{d}\mathcal{J}_w^{\mathrm{opt}} = \mu = -\lambda.}$$

Geometric interpretation

The differential

$$\mathrm{d}\mathcal{J}_w^{\mathrm{opt}} \in T_w^*\mathcal{W}$$

measures the first-order change in the optimal cost produced by an infinitesimal variation of the prescribed task value. The geometric multiplier is exactly this sensitivity covector:

$$\mu = \mathrm{d}\mathcal{J}_w^{\mathrm{opt}}.$$

Thus $\mu(\delta w)$ is the first-order change in the optimal cost associated with the task variation δw. Under the conventional plus-sign Lagrangian, the multiplier is $\lambda = -\mu$, and therefore

$$\mathrm{d}\mathcal{J}_w^{\mathrm{opt}} = -\lambda.$$

Sensitivity needs additional assumptions

The existence of a multiplier at a single constrained minimizer does not by itself guarantee that nearby fibers possess selected minimizers that vary differentiably with the task value. It also does not guarantee that the resulting optimal-value function is differentiable. The sensitivity relation above therefore relies on the stated assumptions of local existence, differentiable selection, and suitable stability of the minimizers.

17 From first-order geometry to numerical updates

The preceding theory characterizes stationary points. To construct an optimization algorithm, we must additionally explain how a tangent descent direction produces a new point on the manifold.

A gradient is a tangent vector:

$$-\operatorname{grad}_g J(v) \in T_v\mathcal{V}.$$

It represents an infinitesimal descent direction, but it is not itself a point of $\mathcal{V}$. On a nonlinear manifold, the coordinate addition of this vector to the point v is not an intrinsic operation. A retraction provides the required map from a tangent displacement back to the manifold.

Definition 17.1 (Retraction)**.** Let $\mathcal{U} \subset T\mathcal{V}$ be an open neighborhood of the zero section of the tangent bundle. A *retraction* on $\mathcal{V}$ is a smooth map

$$R : \mathcal{U} \subset T\mathcal{V} \to \mathcal{V}$$

such that, for every $v \in \mathcal{V}$, the restriction

$$R_v : \mathcal{U}_v \subset T_v\mathcal{V} \to \mathcal{V}, \qquad R_v(\xi) := R(v, \xi),$$

satisfies

$$R_v(0_v) = v, \qquad \mathrm{d}(R_v)_{0_v} = \mathrm{Id}_{T_v\mathcal{V}}\,.$$

Here,

$$\mathcal{U}_v = \{\xi \in T_v\mathcal{V} : (v, \xi) \in \mathcal{U}\}$$

is a neighborhood of 0_v in $T_v\mathcal{V}$, and the canonical identification

$$T_{0_v}(T_v\mathcal{V}) \simeq T_v\mathcal{V}$$

is understood in the second condition.

For a fixed v, the map R_v takes a sufficiently small tangent displacement $\xi \in T_v\mathcal{V}$ and returns a point $R_v(\xi) \in \mathcal{V}$. The first condition states that a zero displacement leaves the point unchanged. The second states that the retraction reproduces the tangent displacement to first order.

More precisely, in a local chart φ around v,

$$\varphi\big(R_v(\xi)\big) = \varphi(v) + [\xi]_\varphi + o\big(\|[\xi]_\varphi\|\big).$$

This coordinate formula describes the first-order behavior of the retraction; it is not its intrinsic definition.

An unconstrained metric-gradient iteration can therefore be written as

$$\boxed{v_{k+1} = R_{v_k}\big(-\alpha_k \,\mathrm{grad}_g J(v_k)\big).}$$

The metric determines the tangent descent direction, while the retraction determines how the finite tangent displacement is realized as a new point of $\mathcal{V}$.

17.1 Optimization on a constraint fiber

Now consider the embedded constraint fiber

$$\mathcal{F}_w = f^{-1}(w) \subset \mathcal{V}.$$

A constrained algorithm must solve two different geometric problems:

1. construct a descent direction belonging to the feasible tangent space $T_v\mathcal{F}_w$;
2. map the resulting tangent displacement to an actual point of $\mathcal{F}_w$.

The first problem is solved by tangent projection. The second is solved by a retraction on the fiber.

Assume that $\mathcal{F}_w$ carries the metric induced by the ambient metric g. Since

$$T_v\mathcal{V} = T_v\mathcal{F}_w \oplus (T_v\mathcal{F}_w)^{\perp_g},$$

every ambient tangent vector $\eta \in T_v\mathcal{V}$ admits a unique decomposition

$$\eta = \eta^{\mathrm{T}} + \eta^{\mathrm{N}},$$

where

$$\eta^{\mathrm{T}} \in T_v\mathcal{F}_w, \qquad \eta^{\mathrm{N}} \in (T_v\mathcal{F}_w)^{\perp_g}.$$

Definition 17.2 (Metric-orthogonal tangent projection)**.** The metric-orthogonal projection onto the feasible tangent space is the linear map

$$\mathrm{Proj}^{g}_{T_v\mathcal{F}_w} : T_v\mathcal{V} \to T_v\mathcal{F}_w$$

defined by

$$\mathrm{Proj}^{g}_{T_v\mathcal{F}_w}(\eta) = \eta^{\mathrm{T}}.$$

Equivalently, it is characterized by

$$\eta - \mathrm{Proj}^{g}_{T_v\mathcal{F}_w}(\eta) \in (T_v\mathcal{F}_w)^{\perp_g}.$$

The projection depends on the metric because the orthogonal complement $(T_v\mathcal{F}_w)^{\perp_g}$ depends on the metric.

Proposition 17.3 (Gradient of the restricted cost)**.** *For the metric induced on $\mathcal{F}_w$,*

$$\boxed{\mathrm{grad}_{\mathcal{F}_w}(J|_{\mathcal{F}_w})(v) = \mathrm{Proj}^{g}_{T_v\mathcal{F}_w}\big(\mathrm{grad}_g\, J(v)\big).}$$

Proof. Let

$$\eta = \mathrm{Proj}^{g}_{T_v\mathcal{F}_w}\big(\mathrm{grad}_g\, J(v)\big).$$

For every $\xi \in T_v\mathcal{F}_w$, the difference $\mathrm{grad}_g\, J(v) - \eta$ is g-orthogonal to ξ. Hence

$$\begin{aligned} g_v(\eta, \xi) &= g_v\big(\mathrm{grad}_g\, J(v), \xi\big) \\ &= \mathrm{d}J_v(\xi) \\ &= \mathrm{d}(J|_{\mathcal{F}_w})_v(\xi). \end{aligned}$$

Thus η satisfies the defining identity of the gradient of the restricted cost. Uniqueness of the gradient proves the result. □

The negative restricted gradient is therefore a feasible descent direction whenever it is nonzero:

$$-\,\mathrm{grad}_{\mathcal{F}_w}(J|_{\mathcal{F}_w})(v) \in T_v\mathcal{F}_w.$$

17.2 Connection with constrained stationarity and multipliers

This projected-gradient formula is the algorithmic counterpart of the multiplier theory developed earlier. Indeed,

$$\mathrm{grad}_{\mathcal{F}_w}(J|_{\mathcal{F}_w})(v) = 0$$

if and only if

$$\mathrm{Proj}^{g}_{T_v\mathcal{F}_w}\big(\mathrm{grad}_g\, J(v)\big) = 0.$$

This means that the ambient gradient has no feasible tangent component and lies entirely in the metric normal space:

$$\mathrm{grad}_g\, J(v) \in (T_v\mathcal{F}_w)^{\perp_g}.$$

Applying the flat map gives the equivalent metric-free statement

$$\mathrm{d}J_v \in (T_v\mathcal{F}_w)^{\circ}.$$

By the annihilator identity, this is equivalent to the existence of the geometric multiplier

$$\mu \in T_w^* \mathcal{W}$$

such that

$$\mathrm{d}J_v = (\mathrm{d}f_v)^* \mu.$$

Consequently,

$$\boxed{\begin{aligned} \operatorname{grad}_{\mathcal{F}_w}(J|_{\mathcal{F}_w})(v) = 0 &\iff \mathrm{d}J_v|_{T_v\mathcal{F}_w} = 0 \\ &\iff \mathrm{d}J_v \in (T_v\mathcal{F}_w)^\circ \\ &\iff \mathrm{d}J_v = (\mathrm{d}f_v)^* \mu. \end{aligned}}$$

Thus the multiplier equation characterizes the zeros of the constrained gradient field. The projected gradient supplies the feasible descent direction used to search for such zeros.

17.3 Finite feasible updates

Tangent projection alone does not produce a new feasible point. It produces a vector in $T_v\mathcal{F}_w$, which is feasible only to first order. A finite ambient addition generally leaves a curved constraint fiber.

Let

$$R^{\mathcal{F}_w} : \mathcal{U}^{\mathcal{F}_w} \subset T\mathcal{F}_w \to \mathcal{F}_w$$

be a retraction on the constraint fiber. Its restriction at $v \in \mathcal{F}_w$ is

$$R_v^{\mathcal{F}_w} : \mathcal{U}_v^{\mathcal{F}_w} \subset T_v\mathcal{F}_w \to \mathcal{F}_w.$$

A constrained metric-gradient iteration is then

$$\boxed{v_{k+1} = R_{v_k}^{\mathcal{F}_w}\left(-\alpha_k \operatorname{grad}_{\mathcal{F}_w}(J|_{\mathcal{F}_w})(v_k)\right).}$$

Equivalently, using the ambient gradient and its tangent projection,

$$\boxed{v_{k+1} = R_{v_k}^{\mathcal{F}_w}\left(-\alpha_k \operatorname{Proj}^g_{T_{v_k}\mathcal{F}_w}\left(\operatorname{grad}_g J(v_k)\right)\right).}$$

Projection and retraction perform different tasks

The metric-orthogonal projection

$$\operatorname{Proj}^g_{T_v\mathcal{F}_w}$$

converts the ambient gradient into a feasible tangent direction. The retraction

$$R_v^{\mathcal{F}_w}$$

converts the resulting tangent displacement into an actual feasible point.
Projection is linear and acts within the tangent space at a fixed point. Retraction is generally nonlinear and maps from the tangent space back to the constraint manifold. Neither operation replaces the other.

18 Unified geometric picture

The complete structure developed in this lecture can be organized into two connected paths in Fig. 3. The upper path starts from the cost differential and leads, through a metric, to an unconstrained numerical update. The lower path introduces a task fiber, restricts the admissible tangent directions, and leads both to the multiplier characterization of stationarity and to a constrained numerical update.

19 Coordinate-free statements and coordinate implementation

The following two tables summarize how the intrinsic constructions developed in this lecture are represented in coordinates. Part I covers the differential, metric gradients, and metric selection. Part II covers constraint fibers, multipliers, and numerical updates on the constraint manifold.

Part I: Differential, gradients, and metric selection

Object or operation	Intrinsic statement	Coordinate representation
Cost differential and metric gradient		
Cost differential	$\mathrm{d}J_v \in T_v^*\mathcal{V}$	$d_v = \big(\partial J/\partial x^i(v)\big)_i$
Level directions	$\ker \mathrm{d}J_v \subset T_v\mathcal{V}$	$d_v^\top [\xi]_v = 0$
Descent test	$\mathrm{d}J_v(\xi) < 0$	$d_v^\top [\xi]_v < 0$
Metric gradient	$\mathrm{grad}_g\, J(v) = \sharp_{g,v}(\mathrm{d}J_v)$	$[\mathrm{grad}_g\, J]_v = G_v^{-1} d_v$
Steepest descent direction	$\arg \min\limits_{\Vert\xi\Vert_g=1} \mathrm{d}J_v(\xi)$	$\arg \min\limits_{\xi^\top G_v \xi=1} d_v^\top \xi$
Regularized gradient step	$\arg\min\limits_{\xi} \left[\mathrm{d}J_v(\xi) + \frac{1}{2\eta}\Vert\xi\Vert_g^2\right]$	$-\eta G_v^{-1} d_v$
Coordinate change	Vectors and covectors transform so that $\mathrm{d}J_v(\xi)$ is invariant	$[\xi]_y = K[\xi]_x, \quad [\mathrm{d}J]_y = K^{-\top}[\mathrm{d}J]_x$
Metric selection, preconditioning, and Newton		
Preconditioned direction	Negative sharp image under a selected metric	$-P_v d_v$, with $P_v = G_v^{-1} \succ 0$
Newton direction	Negative gradient under the Hessian metric, when $H_v \succ 0$	$-H_v^{-1} d_v$
Stationary-point invariance	$\mathrm{d}J_v = 0 \Longleftrightarrow \mathrm{grad}_g\, J(v) = 0$	$d_v = 0 \Longleftrightarrow G_v^{-1} d_v = 0$

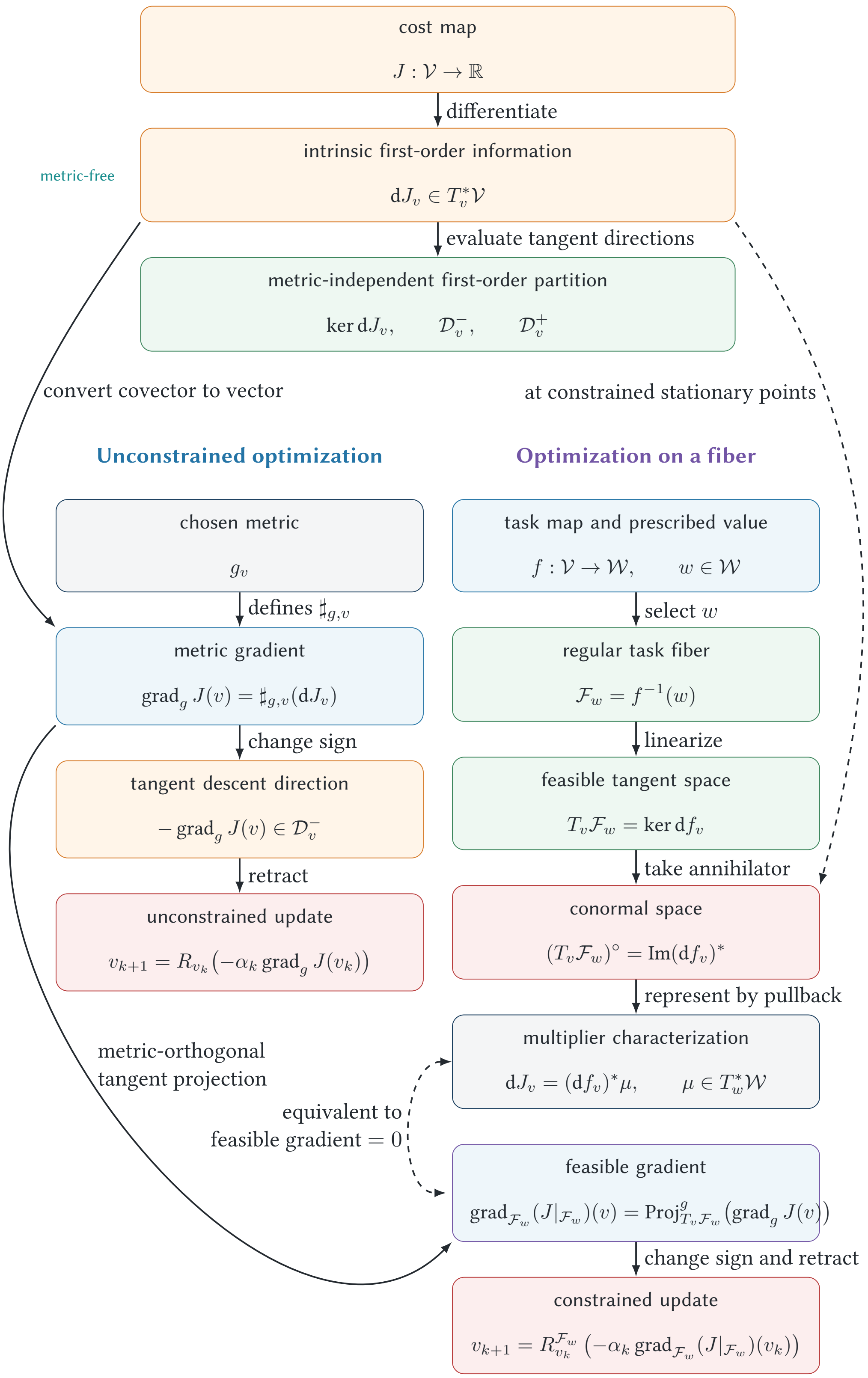


Figure 3: The cost first supplies the metric-independent differential $\mathrm{d}J_v$, which determines the level, ascent, and descent directions. The left column shows how a metric converts this covector into an unconstrained gradient direction and how a retraction produces a new point of $\mathcal{V}$. The right column shows how the task map defines the constraint fiber, its feasible tangent space, and its conormal space. The multiplier equation characterizes constrained stationary points, while the projected gradient and a fiber retraction provide a feasible numerical update.

Part II: Constraint geometry, multipliers, and feasible updates

Object or operation	Intrinsic statement	Coordinate representation
Constraint fibers, conormals, and multipliers		
Task fiber	$\mathcal{F}_w = f^{-1}(w)$	$f(v) = w$
Feasible tangent space	$T_v\mathcal{F}_w = \ker \mathrm{d}f_v$	$F_v[\xi]_v = 0$
Restricted differential	$\mathrm{d}(J\vert_{\mathcal{F}_w})_v = (\mathrm{d}\iota_{w,v})^*(\mathrm{d}J_v)$	Restrict $d_v^\top[\xi]_v$ to vectors satisfying $F_v[\xi]_v = 0$
Constrained stationarity	$\mathrm{d}J_v\vert_{T_v\mathcal{F}_w} = 0$	$d_v^\top[\xi]_v = 0$ for every $[\xi]_v \in \ker F_v$
Conormal space	$(T_v\mathcal{F}_w)^\circ = \mathrm{Im}(\mathrm{d}f_v)^*$	$\mathrm{Im}\, F_v^\top$
Geometric multiplier	$\mathrm{d}J_v = (\mathrm{d}f_v)^*\mu$	$d_v = F_v^\top[\mu]_w$
Plus-sign multiplier	$\mathrm{d}J_v + (\mathrm{d}f_v)^*\lambda = 0$, $\lambda = -\mu$	$d_v + F_v^\top[\lambda]_w = 0$
Conormal-to-normal conversion	$\sharp_{g,v}\big((T_v\mathcal{F}_w)^\circ\big) = (T_v\mathcal{F}_w)^{\perp_g}$	A conormal column a_v becomes $G_v^{-1}a_v$
Optimal-value sensitivity	$\mathrm{d}\mathcal{J}_w^{\mathrm{opt}} = \mu = -\lambda$	$\delta\mathcal{J}^{\mathrm{opt}} \approx [\mu]_w^\top[\delta w]_w$
Numerical updates on manifolds		
Restricted gradient	$\mathrm{grad}_{\mathcal{F}_w}(J\vert_{\mathcal{F}_w})(v) = \mathrm{Proj}^g_{T_v\mathcal{F}_w}\big(\mathrm{grad}_g\, J(v)\big)$	Project the ambient gradient components onto $\ker F_v$ using G_v
Unconstrained update	$v_{k+1} = R_{v_k}\big(-\alpha_k\, \mathrm{grad}_g\, J(v_k)\big)$	Compute the tangent components, then evaluate R_{v_k}
Constrained update	$v_{k+1} = R^{\mathcal{F}_w}_{v_k}\big(-\alpha_k\, \mathrm{grad}_{\mathcal{F}_w}(J\vert_{\mathcal{F}_w})(v_k)\big)$	Project onto $\ker F_{v_k}$, then evaluate the fiber retraction

Final synthesis

Takeaway

The cost

$$J : \mathcal{V} \to \mathbb{R}$$

supplies the intrinsic first-order covector

$$\mathrm{d}J_v \in T_v^*\mathcal{V}.$$

This differential, rather than any particular coordinate formula or gradient vector, is the metric-independent starting point of first-order optimization. It determines the tangent directions of zero first-order change and separates the intrinsic ascent and descent half-spaces.
A Riemannian metric converts the differential into a tangent vector:

$$\mathrm{grad}_g\, J(v) = \sharp_{g,v}(\mathrm{d}J_v).$$

The metric determines lengths, orthogonality, normal directions, and the meaning of steepest descent. Different metrics generally produce different gradients from the same differential, while the stationary set and the intrinsic ascent/descent partition remain unchanged. Coordinates represent the differential, metric, and gradient by compatible arrays, leading to

$$[\operatorname{grad}_g J]_v = G_v^{-1} d_v.$$

Preconditioning is the coordinate realization of choosing an inverse metric. When the Hessian is positive definite, Newton's direction is the negative gradient associated with the Hessian-based metric. Newton therefore uses second-order information to select the geometry in which the differential is converted into a descent vector.
An equality constraint

$$f(v) = w$$

selects the regular fiber

$$\mathcal{F}_w = f^{-1}(w), \qquad T_v\mathcal{F}_w = \ker \mathrm{d}f_v.$$

Constrained stationarity requires the cost differential to vanish on this feasible tangent space:

$$\mathrm{d}J_v|_{T_v\mathcal{F}_w} = 0.$$

Equivalently, the cost differential belongs to the conormal space and is the pullback of a task-space multiplier covector:

$$\mathrm{d}J_v \in (T_v\mathcal{F}_w)^\circ = \mathrm{Im}(\mathrm{d}f_v)^*, \qquad \mathrm{d}J_v = (\mathrm{d}f_v)^*\mu.$$

The familiar Jacobian transpose is the coordinate representation of this pullback. Under the stated sensitivity assumptions, the same multiplier is the differential of the optimal-value function:

$$\mathrm{d}\mathcal{J}_w^{\mathrm{opt}} = \mu = -\lambda.$$

Finally, first-order geometry supplies tangent directions, not finite manifold displacements. In constrained optimization, metric projection converts the ambient gradient into a feasible tangent direction, while a retraction converts that tangent displacement into a new feasible point. Gradient, preconditioning, Newton, multipliers, projection, and retraction are therefore not disconnected formulas. They are complementary realizations of one intrinsic first-order geometric structure centered on the differential.

Some Exercises

The following exercises follow the conceptual progression of the lecture, from intrinsic differentials and coordinate representations to metric selection, constrained stationarity, and feasible numerical updates.

Exercise 19.1 (One function, two coordinate representations)**.** Consider the one-dimensional manifold $\mathcal{V} = \mathbb{R}$. Suppose that the cost is represented in the coordinate x by

$$J_x(x) = x^2.$$

Introduce the coordinate

$$y = \arctan x.$$

(a) Determine the coordinate representative $J_y(y)$.

(b) Compute the coordinate expressions of $\mathrm{d}J$ in the x- and y-coordinates.

(c) Use the transformation relation between $\mathrm{d}x$ and $\mathrm{d}y$ to verify that the two expressions represent the same intrinsic one-form.

Exercise 19.2 (The level hyperplanes of a covector)**.** Let $\alpha \in (\mathbb{R}^2)^*$ be defined by

$$\alpha(\xi_1, \xi_2) = \xi_1 + 2\xi_2.$$

(a) Determine the kernel $\ker \alpha$.

(b) For $a \in \mathbb{R}$, describe the affine level set

$$H_a = \{\xi \in \mathbb{R}^2 : \alpha(\xi) = a\}.$$

(c) Identify the positive and negative half-spaces determined by α.

(d) Compare the labelled level hyperplanes of α and 2α. State what remains unchanged and what additional information is lost if only the kernel is retained.

Exercise 19.3 (One differential and two metric gradients)**.** On $\mathcal{V} = \mathbb{R}^2$, consider

$$J(z) = \frac{1}{2}(9z_1^2 + z_2^2)$$

and the two constant metrics represented in the selected coordinates by

$$G_1 = I, \qquad G_2 = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}.$$

(a) Compute the component column d_z of $\mathrm{d}J_z$.

(b) Compute $[\operatorname{grad}_{g_1} J]_z$ and $[\operatorname{grad}_{g_2} J]_z$.

(c) Verify directly through the intrinsic pairing test that both negative gradients are strict descent directions whenever $z \neq 0$.

(d) Explain which first-order objects remain unchanged when the metric is changed and which objects depend on the metric.

Exercise 19.4 (Coordinate covariance in polar coordinates)**.** On $\mathbb{R}^2 \setminus \{0\}$, consider

$$J(x, y) = xy.$$

Introduce polar coordinates through

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

(a) Determine the coordinate representative $J_{r\theta}$.

(b) Compute the components of $\mathrm{d}J$ in Cartesian and polar coordinates.

(c) Transform the Cartesian covector components using the inverse-transpose Jacobian and verify agreement with the direct polar-coordinate computation.

(d) Represent the standard metric in polar coordinates and compute the corresponding gradient components.

(e) Verify that the gradient components transform as those of a tangent vector.

(f) Explain why replacing the polar metric matrix by I would select a different metric rather than merely a different representation of the same metric.

Exercise 19.5 (Hessian metric and Newton direction)**.** Let

$$J(z) = \frac{1}{2}z^\top Qz, \qquad Q = Q^\top \succ 0.$$

(a) Compute d_z and the Hessian H_z.

(b) Compute the negative gradient direction for the metric represented by $G_1 = I$.

(c) Compute the negative gradient direction for the Hessian-based metric represented by $G_2 = Q$.

(d) Show that the second direction coincides with the Newton direction.

(e) Show that a unit step along the Newton direction reaches the minimizer in one iteration.

(f) Explain geometrically how the Hessian metric compensates for the anisotropy of the quadratic cost.

(g) Explain why this interpretation does not imply that Newton's method is a first-order method.

Exercise 19.6 (From feasible tangents to the multiplier)**.** Let

$$f : \mathcal{V} \to \mathcal{W}$$

be smooth, let $w \in \mathcal{W}$ be a regular value, and let

$$\mathcal{F}_w = f^{-1}(w).$$

Suppose that $v \in \mathcal{F}_w$ is stationary for $J|_{\mathcal{F}_w}$.

(a) Starting from

$$T_v\mathcal{F}_w = \ker \mathrm{d}f_v,$$

show that

$$\mathrm{d}J_v \in (\ker \mathrm{d}f_v)^\circ.$$

(b) Use the annihilator identity

$$(\ker L)^\circ = \mathrm{Im}\, L^*$$

to prove that there exists $\mu \in T_w^*\mathcal{W}$ satisfying

$$\mathrm{d}J_v = (\mathrm{d}f_v)^*\mu.$$

(c) Explain precisely why regularity of the fiber makes the multiplier μ unique.

(d) Write the same condition using the conventional plus-sign Lagrangian multiplier λ.

Exercise 19.7 (Generalized Rayleigh-quotient problem)**.** Let

$$J(v) = \frac{1}{2}v^\top Av, \qquad f(v) = \frac{1}{2}v^\top Bv,$$

where

$$A = A^\top, \qquad B = B^\top \succ 0,$$

and consider the fiber corresponding to $w = 1/2$.

(a) Compute $\mathrm{d}J_v$ and $\mathrm{d}f_v$.

(b) Derive the intrinsic multiplier equation

$$\mathrm{d}J_v = \mu\,\mathrm{d}f_v.$$

(c) Use the selected inner product to obtain the generalized eigenvalue equation

$$Av = \mu Bv.$$

(d) Define the generalized Rayleigh quotient

$$\mathcal{R}_{A,B}(v) = \frac{v^\top Av}{v^\top Bv}, \qquad v \neq 0,$$

and show that its stationary value at a generalized eigenvector is the corresponding generalized eigenvalue.

(e) Characterize the constrained stationary points on the fiber $v^\top Bv = 1$.

Exercise 19.8 (Projection and retraction on the sphere)**.** Let

$$\mathcal{F}_{1/2} = \mathbb{S}^{n-1} \subset \mathbb{R}^n,$$

and let $v \in \mathbb{S}^{n-1}$.

(a) Show that

$$T_v\mathbb{S}^{n-1} = \{\xi \in \mathbb{R}^n : v^\top \xi = 0\}.$$

(b) For an ambient vector $\eta \in \mathbb{R}^n$, derive its orthogonal projection onto $T_v\mathbb{S}^{n-1}$.

(c) Show that, for a nonzero $\xi \in T_v\mathbb{S}^{n-1}$, the finite ambient addition $v + \xi$ does not generally belong to the sphere.

(d) Consider the normalization map

$$R_v(\xi) = \frac{v + \xi}{\|v + \xi\|}.$$

Verify that it maps sufficiently small tangent displacements back to the sphere.

(e) Verify the two defining retraction properties at $\xi = 0_v$.

(f) Explain the distinct roles played by tangent projection and retraction in one constrained gradient iteration.

Exercise 19.9 (Optional: sequential positive-definite maps)**.** Let

$$d = \begin{bmatrix} 10 \\ 1 \end{bmatrix}, \qquad P_1 = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}.$$

(a) Verify that P_1 and P_2 are symmetric positive definite.

(b) Show that

$$-P_1 d$$

is a descent direction for the differential represented by d.

(c) Compute

$$-P_2 P_1 d$$

and show that it is an ascent direction.

(d) Explain why interpreting both P_1 and P_2 as successive inverse-metric maps is not type-correct.

(e) State the different interpretation required to regard P_2 as acting on the vector produced by P_1.