DUAL CONNECTIONS IN INFORMATION GEOMETRY

From divergences to optimization in ML

Samuele Mongodi — UniMiB

INFORMATION GEOMETRY

Part 1 — Why geometry? Metric, connections, duality

Roadmap: statistical models → Fisher metric → affine connections → dual connections → divergence generates all.

WHY INTRODUCE A GEOMETRIC LANGUAGE?

- In ML we optimize over **parameters**, but what we really change is a **distribution** (or a predictive model) $p(\cdot;\theta)$.
- The same model can be described by many parametrizations $\theta \mapsto \tilde{\theta}(\theta)$: algorithms should behave invariantly under reparametrization.
- Euclidean geometry in parameter space is often arbitrary: it depends on the chosen coordinates.
- Geometry provides intrinsic notions of **length**, **angle**, **straightness** on the space of models.

LOSS FUNCTIONS ARE OFTEN DIVERGENCES

- Many objectives compare distributions: cross-entropy / KL, negative log-likelihood, f-divergences, Bregman-type losses.
- A divergence $D(p\|q)$ is typically **asymmetric**, but it encodes meaningful local geometry.
- The key IG idea: a single divergence can induce a metric g and a pair of dual affine connections (∇, ∇^*).

Interpretation: "move in parameter space" $\leftrightarrow$ "move on a curved manifold of distributions".

WHY DO WE NEED CONNECTIONS (NOT ONLY A METRIC)?

- A metric g gives inner products and distances, but learning also needs a notion of **straight paths** and **interpolation**.
- In statistics there are (at least) two natural interpolations between models: mixtures and exponential tilting.
- These lead to different "geodesics" → captured by **affine connections**.
- Dual connections formalize the coexistence of two compatible straightness notions.

STATISTICAL MODEL AS A MANIFOLD

- A model is a family $\mathcal{M} = \{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}^d\}$.
- View Θ as coordinates on a manifold M (dimension d): each θ labels a distribution p_θ.
- Reparametrizations are changes of coordinates on M.

Invariant viewpoint: algorithms should depend on p_θ, not on the particular chart θ.

TANGENT VECTORS AND THE SCORE

- Infinitesimal change: $\delta p \approx \sum_i \delta\theta^i \, \partial_i p$.
- Use the score functions

$$\partial_i \log p(x;\theta) = \frac{\partial_i p(x;\theta)}{p(x;\theta)}.$$

- Under regularity, $\mathbb{E}_\theta[\partial_i \log p] = 0$, so scores behave like centered features.

FISHER INFORMATION METRIC

- Define the Riemannian metric $g_{ij}(\theta) = \mathbb{E}_\theta[\partial_i \log p \; \partial_j \log p]$.
- Equivalent expression (regular case): $g_{ij}(\theta) = -\mathbb{E}_\theta[\partial_i \partial_j \log p]$.
- Interpretation: local distinguishability / sensitivity of the model.
- Crucial property: **coordinate invariance**.

Preview: steepest descent w.r.t. g = *natural gradient* (Part 4).

METRIC FROM A DIVERGENCE (2ND ORDER TERM)

- Let $D(\theta \| \theta') := D(p_\theta \| p_{\theta'})$ with $D(\theta \| \theta) = 0$.
- Local expansion around $\theta' = \theta$:

$$D(\theta \| \theta + d\theta) = \tfrac{1}{2}\, g_{ij}(\theta)\, d\theta^i d\theta^j + O(\|d\theta\|^3).$$

- So g is encoded in the second derivatives of D on the diagonal.

AFFINE CONNECTIONS: "STRAIGHTNESS" ON M

- An affine connection ∇ defines covariant derivatives $\nabla_X Y$ and geodesics (auto-parallels).
- Different connections correspond to different straight paths between models (mixture vs exponential interpolation).
- In IG we keep track of two compatible straightness notions → dual connections.

DUAL CONNECTIONS (∇, ∇^*)

- ∇ and ∇^* are dual w.r.t. g if

$$X\, g(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

- This is a generalized metric-compatibility: ∇ and ∇^* "share" the same metric.
- Levi–Civita is the self-dual special case: $\nabla = \nabla^*$.

DIVERGENCE ⟶ METRIC (DERIVATIVE RECIPE)

- Mixed second derivatives give the metric:

$$g_{ij}(\theta) = \frac{\partial^2}{\partial\theta^i\, \partial\theta'^j} D(\theta \| \theta') \Big|_{\theta'=\theta}.$$

- This matches the 2nd-order expansion of D near the diagonal.

DIVERGENCE ⟶ DUAL CONNECTIONS (3RD DERIVATIVES)

- Third derivatives define Christoffel symbols (schematically):

$$\Gamma_{ijk}(\theta) \;\sim\; -\partial_i \partial_j \partial'_k \, D(\theta \| \theta')\big|_{\theta'=\theta},$$

$$\Gamma^*_{ijk}(\theta) \;\sim\; -\partial'_i \partial'_j \partial_k \, D(\theta \| \theta')\big|_{\theta'=\theta}.$$

- One divergence induces the triple $\left(g, \nabla, \nabla^*\right)$.
- Canonical example: KL divergence $\Rightarrow$ Fisher metric + the (e,m) dual structure.

PART 1 RECAP

- We want invariance: optimize on the manifold of distributions, not in arbitrary coordinates.
- A divergence gives a local metric (2nd order) and dual connections (3rd order).
- Dual connections encode two natural straightness notions in statistics.

Next: α-connections, dually flat manifolds, Legendre duality, Bregman divergences (Part 2).

PART 2 — DUAL AFFINE GEOMETRY

α-connections · Dually flat manifolds · Legendre duality · Bregman divergence

Goal: see how dual connections become concrete via α-connections, and how flatness leads to Legendre/Bregman geometry.

α-CONNECTIONS: A ONE-PARAMETER FAMILY

- In information geometry there is not a single "natural" connection: different statistical operations suggest different notions of straightness.
- Amari's α-connections $\nabla^{(\alpha)}$ interpolate between two extremes.
- They are torsion-free affine connections compatible with the Fisher metric in the dual sense.
- Duality relation: $\left(\nabla^{(\alpha)}\right)^* = \nabla^{(-\alpha)}$.

TWO KEY CASES: EXPONENTIAL VS MIXTURE

- $\alpha = +1$: the exponential connection (often called e-connection), denoted $\nabla^{(1)}$.
- $\alpha = -1$: the mixture connection (often called m-connection), denoted $\nabla^{(-1)}$.
- Duality: $(\nabla^{(1)})^* = \nabla^{(-1)}$ and vice versa.
- $\alpha = 0$ gives the Levi–Civita connection of the Fisher metric.

Statistical meaning: "straight" paths can correspond to *mixing* or *exponential tilting*.

WHAT ARE THE E- AND M-CONNECTIONS (EXPLICITLY)?

- Let $\ell(x;\theta) = \log p(x;\theta)$ and $g_{ij}(\theta) = \mathbb{E}_\theta[\partial_i \ell \, \partial_j \ell]$.
- The α-connection coefficients (lowered index form) can be written as

$$\Gamma^{(\alpha)}_{ijk}(\theta) = \mathbb{E}_\theta\big[\partial_i\partial_j\ell \, \partial_k\ell\big] + \frac{1-\alpha}{2}\,\mathbb{E}_\theta\big[\partial_i\ell \, \partial_j\ell \, \partial_k\ell\big].$$

- e-connection ($\alpha = +1$):

$$\Gamma^{(1)}_{ijk}(\theta) = \mathbb{E}_\theta\big[\partial_i\partial_j\ell \, \partial_k\ell\big].$$

- m-connection ($\alpha = -1$):

$$\Gamma^{(-1)}_{ijk}(\theta) = \mathbb{E}_\theta\big[\partial_i\partial_j\ell \, \partial_k\ell\big] + \mathbb{E}_\theta\big[\partial_i\ell \, \partial_j\ell \, \partial_k\ell\big].$$

- Duality: $\left(\nabla^{(1)}\right)^* = \nabla^{(-1)}$ w.r.t. the Fisher metric.

WHAT DO E AND M MEAN GEOMETRICALLY?

- A connection ∇ is flat if it admits affine coordinates: in those coordinates, ∇-geodesics are ordinary straight lines.
- e-connection (exponential geometry): there exist coordinates θ such that, for exponential families,

$$\log p(x;\theta) = \langle \theta,\ F(x) \rangle - \psi(\theta) + c(x),$$

  i.e. $\log p(\cdot;\theta)$ is affine in θ (up to the normalization term ψ).
- m-connection (mixture geometry): there exist dual coordinates η (expectation / mixture parameters) in which convex combinations are affine:

$$p_t = (1-t)\,p_0 + t\,p_1 \quad \Longleftrightarrow \quad \eta(t) = (1-t)\eta_0 + t\eta_1.$$

- **Practical slogan:** e-straight = "straight in natural parameters / log-density"; m-straight = "straight in mixture / expectation parameters".
- These two straightness notions are dual (and will become fully explicit in *Part 3*).

INTUITION: TWO NOTIONS OF "LINE" BETWEEN DISTRIBUTIONS

- **Mixture path** (linear in densities):

$$p_t = (1-t)p_0 + tp_1.$$

- **Exponential path** (linear in log-densities, up to normalization):

$$\log p_t \ \propto\ (1-t)\log p_0 + t \log p_1.$$

- These correspond, in the right coordinate systems, to geodesics of dual connections.

DUALLY FLAT MANIFOLDS

- A statistical manifold (M, g, ∇, ∇^*) is dually flat if both ∇ and ∇^* are flat (curvature $= 0$).
- Flatness means we can find global (or large-chart) affine coordinates: θ affine for ∇, and η affine for ∇^*.
- In these coordinates, the corresponding geodesics are "straight lines" ($\theta(t)$ linear or $\eta(t)$ linear).

Payoff: dual flatness gives a convex-analytic description via potentials and Legendre duality.

POTENTIALS AND LEGENDRE DUALITY

- In the dually flat case there exist convex potentials $\psi(\theta)$ and $\varphi(\eta)$.
- Dual coordinates are linked by

$$\eta = \nabla_\theta \psi(\theta), \qquad \theta = \nabla_\eta \varphi(\eta).$$

- φ is the Legendre transform of ψ:

$$\varphi(\eta) = \sup_\theta \big(\langle \theta, \eta \rangle - \psi(\theta)\big).$$

- The Fisher metric becomes a Hessian metric:

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta).$$

BREGMAN DIVERGENCE (CANONICAL DIVERGENCE IN FLAT COORDINATES)

- Given a strictly convex differentiable ψ, the Bregman divergence is

$$D_\psi(\theta \| \theta') = \psi(\theta) - \psi(\theta') - \langle \nabla\psi(\theta'), \theta - \theta' \rangle.$$

- It is generally asymmetric, but encodes a clear geometric meaning: "gap" between ψ and its supporting hyperplane at θ'.
- In dually flat IG, the canonical divergence can be written as a Bregman divergence (in appropriate coordinates).

PART 2 RECAP

- α-connections provide a continuum of affine structures; duality is $\alpha \leftrightarrow -\alpha$.
- Dually flat $\Rightarrow$ dual affine coordinates (θ, η) and convex potentials.
- Legendre duality organizes the geometry; the divergence becomes Bregman-like.
- Geodesics become simple in the right coordinates.

Next: exponential families as the canonical dually flat example (Part 3).

PART 3 — WORKED EXAMPLE

Exponential families as a dually flat statistical manifold

We will compute explicitly: manifold · Fisher metric · KL divergence · dual coordinates · (e,m)-flatness.

EXPONENTIAL FAMILY: THE PROBABILISTIC SETTING

- Fix a base measure μ and sufficient statistics $F(x) = (F_1(x), \ldots, F_d(x))$.
- Define a family of densities (w.r.t. μ):

$$p(x;\theta) = \exp\big(\langle \theta, F(x)\rangle - \psi(\theta)\big)\, h(x), \qquad \theta \in \Theta \subset \mathbb{R}^d.$$

- $\psi(\theta)$ is the log-partition function, chosen so that $\int p(x;\theta)\, d\mu = 1$:

$$\psi(\theta) = \log \int \exp(\langle \theta, F(x)\rangle)\, h(x)\, d\mu(x).$$

- The parameter space Θ is (typically) an open convex set where ψ is finite and smooth.

MANIFOLD OF DISTRIBUTIONS

- The model manifold is $M = \{p_\theta : \theta \in \Theta\}$, with chart θ (e-coordinates).
- The log-density is affine in θ (up to $-\psi(\theta)$):

$$\log p(x;\theta) = \langle \theta, F(x) \rangle - \psi(\theta) + \log h(x).$$

- This affine structure is exactly what makes exponential families e-flat.

DUAL COORDINATES: EXPECTATION PARAMETERS

- Define the m-coordinates (expectation parameters)

$$\eta(\theta) = \mathbb{E}_\theta[F(X)] \in \mathbb{R}^d.$$

- Compute η from ψ:

$$\eta_i(\theta) = \partial_i \psi(\theta).$$

- Under regularity/minimality, $\theta \mapsto \eta$ is a diffeomorphism onto its image.
- So the same point $p \in M$ can be coordinatized as θ or η.

Key relation: $\eta = \nabla_\theta \psi(\theta)$ (Legendre-type duality appears naturally).

FISHER METRIC: EXPLICIT FORMULA

- The score is $\partial_i \log p(x;\theta) = F_i(x) - \partial_i \psi(\theta) = F_i(x) - \eta_i$.
- Hence the Fisher information is a covariance matrix:

$$g_{ij}(\theta) = \mathbb{E}_\theta[(F_i - \eta_i)(F_j - \eta_j)] = \mathrm{Cov}_\theta(F_i, F_j).$$

- Also,

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta),$$

so g is a Hessian metric with potential ψ.

MINI-DERIVATION: KL = BREGMAN (EXPONENTIAL FAMILY)

- Start from $\log p(x;\theta) = \langle \theta, F(x) \rangle - \psi(\theta) + \log h(x)$.
- Compute the log-ratio:

$$\log \frac{p(x;\theta)}{p(x;\theta')} = \langle \theta - \theta', F(x) \rangle - \big(\psi(\theta) - \psi(\theta')\big).$$

- Take expectation under p_θ and use $\mathbb{E}_\theta[F] = \nabla\psi(\theta)$:

$$\mathrm{KL}(p_\theta \| p_{\theta'}) = \langle \theta - \theta', \nabla\psi(\theta) \rangle - \big(\psi(\theta) - \psi(\theta')\big).$$

- Rearrange: $\mathrm{KL}(p_\theta \| p_{\theta'}) = \psi(\theta') - \psi(\theta) - \langle \nabla\psi(\theta), \theta' - \theta \rangle = D_\psi(\theta' \| \theta)$.

Conclusion: KL induces a Hessian metric $g = \nabla^2\psi$ and the dual flat (e, m) geometry.

LEGENDRE DUAL POTENTIAL AND DUAL METRIC

- Define the Legendre dual (convex conjugate)

$$\varphi(\eta) = \sup_{\theta} \big(\langle \theta, \eta \rangle - \psi(\theta)\big).$$

- Duality relations:

$$\eta = \nabla \psi(\theta), \qquad \theta = \nabla \varphi(\eta).$$

- The Fisher metric in η-coordinates is also Hessian:

$$g^{ij}(\eta) = \partial^i \partial^j \varphi(\eta),$$

  i.e. φ controls the dual geometry.

DUAL CONNECTIONS AND FLATNESS (CONCRETE STATEMENT)

- Exponential families are the canonical example of a dually flat manifold: the e-connection and m-connection are both flat.
- In θ (natural parameters), the e-connection is affine: its Christoffel symbols vanish in these coordinates (so e-geodesics are straight in θ).
- In η (expectation parameters), the m-connection is affine: its Christoffel symbols vanish in these coordinates (so m-geodesics are straight in η).
- The two are dual w.r.t. the Fisher metric: $(\nabla^{(1)})^* = \nabla^{(-1)}$.

PART 3 RECAP

- Exponential families give an explicit manifold $M = \{p_\theta\}$ with dual coordinates $\theta \leftrightarrow \eta$.
- Fisher metric: $g = \nabla^2 \psi$ (covariance of sufficient statistics).
- KL divergence is Bregman: $\mathrm{KL}(p_\theta \| p_{\theta'}) = D_\psi(\theta' \| \theta)$.
- Therefore the geometry is dually flat: e-straight = linear in θ, m-straight = linear in η.

Next: how this feeds *natural gradient*, *mirror descent*, and ML algorithms (Part 4).

PART 4 — OPTIMIZATION VIEWPOINTS

Natural gradient · Mirror descent · Dually flat equivalences

Goal: show how IG turns “choose a step” into a geometric choice driven by the loss/divergence.

WHY EUCLIDEAN GRADIENT CAN BE THE WRONG GEOMETRY

- Standard gradient descent uses the Euclidean norm in parameter space: $\theta_{t+1} = \theta_t - \epsilon \nabla f(\theta_t)$.
- But the same model distribution can be described by many parametrizations $\theta \mapsto \tilde{\theta}(\theta)$.
- Euclidean steps depend strongly on coordinates; they do not reflect the "true" change in the model p_θ.
- Information geometry suggests: measure step size using a divergence (locally: Fisher metric).

NATURAL GRADIENT = STEEPEST DESCENT W.R.T. FISHER METRIC

- On a Riemannian manifold (M, g), the steepest descent direction solves

$$\arg \min_{\|\delta\|_g = 1} \langle \nabla f, \delta \rangle.$$

- In coordinates, the Riemannian gradient is

$$\nabla^{\text{nat}} f(\theta) = g(\theta)^{-1} \nabla f(\theta).$$

- For statistical models, $g(\theta)$ is the Fisher information $F(\theta)$.
- Update rule:

$$\theta_{t+1} = \theta_t - \epsilon F(\theta_t)^{-1} \nabla f(\theta_t).$$

Interpretation: move "as much as possible" in decreasing f while keeping model-change small.

NATURAL GRADIENT FROM A KL TRUST REGION

- Consider a local constrained step:

$$\min_{\delta} \ \nabla f(\theta)^{\top}\delta \quad \text{s.t.} \quad \mathrm{KL}(p_\theta \| p_{\theta+\delta}) \le \varepsilon.$$

- Use the 2nd-order expansion $\mathrm{KL}(p_\theta \| p_{\theta+\delta}) \approx \frac{1}{2}\,\delta^{\top} F(\theta)\,\delta$.
- Lagrange multipliers give

$$\delta^{\star} \propto -F(\theta)^{-1}\nabla f(\theta).$$

- So natural gradient is the first-order solution of "minimize f with bounded information-distance step".

PRACTICAL COMPUTATION: WHAT CHANGES VS VANILLA GD

- Vanilla GD needs ∇f. Natural gradient needs (approximately) solving $F(\theta)\,v = \nabla f(\theta)$.
- You rarely invert F explicitly: solve linear systems (e.g. CG), or use structured approximations.
- In probabilistic models, F can be estimated from minibatches using score outer products.
- In deep learning, common approximations include diagonal/low-rank/block-diagonal (details in Part 5).

REMINDER: MIRROR DESCENT (MD)

- MD replaces Euclidean proximity by a Bregman divergence D_ψ from a convex potential ψ.
- Generic update (unconstrained):

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \langle \nabla f(\theta_t),\, \theta - \theta_t \rangle + \frac{1}{\epsilon}\, D_\psi(\theta \| \theta_t) \right\}.$$

- ψ encodes the geometry: Euclidean GD is recovered by $\psi(\theta) = \frac{1}{2}\|\theta\|^2$.
- MD is especially effective when the domain has structure (simplex, PSD cone, sparsity, etc.).

MD IN DUAL COORDINATES (LEGENDRE TRANSFORM)

- Define dual coordinates $\eta = \nabla\psi(\theta)$ (as in dually flat geometry).
- The MD update becomes a simple additive step in η:

$$\eta_{t+1} = \eta_t - \epsilon\, \nabla f(\theta_t), \qquad \theta_{t+1} = \nabla\varphi(\eta_{t+1}),$$

  where φ is the convex conjugate of ψ.
- So MD is "gradient descent in the dual space" + mapping back via Legendre duality.

Bridge to IG: (θ, η) are exactly the dual affine coordinates in dually flat manifolds.

NATURAL GRADIENT ↔ MIRROR DESCENT IN DUALLY FLAT GEOMETRY

- In a dually flat manifold, the Fisher metric is Hessian:

$$g(\theta) = \nabla^2 \psi(\theta).$$

- A first-order step measured with the Bregman divergence D_ψ leads to mirror descent.
- Locally, $D_\psi(\theta + \delta \| \theta) \approx \frac{1}{2}\, \delta^\top g(\theta)\, \delta$, i.e. the Bregman geometry matches the Riemannian metric to second order.
- Therefore, in dually flat models (e.g. exponential families), MD and NG can be seen as two faces of the same geometry: "choose step size via divergence" vs "steepest descent via Fisher metric".

WHERE DOES THIS SHOW UP IN ML PRACTICE?

- Natural gradient: invariance to reparametrization; widely used in probabilistic models and policy optimization.
- Mirror descent: foundational in convex optimization; appears as multiplicative weights / exponentiated gradient.
- Common theme: the divergence determines the geometry and often the update form.

PART 4 RECAP

- Natural gradient = steepest descent under Fisher metric = first-order KL trust-region step.
- Mirror descent = gradient step measured with a Bregman divergence induced by a convex potential.
- In dually flat models, (θ, η) coordinates + Legendre duality unify these views.

Next: broader ML applications (2nd-order approximations, preconditioning, geometry vs loss) — Part 5.

PART 5 — BEYOND THE BASICS

2nd-order optimization · Preconditioning · Choosing the "right" geometry

Thesis: the loss/divergence you care about suggests a geometry; dual connections explain the "two-sided" structure behind many updates.

SECOND-ORDER THINKING: WHY PRECONDITIONING HELPS

- Gradient descent uses a single global scale ϵ, but different directions can have very different sensitivity.
- A generic preconditioned update looks like

$$\theta_{t+1} = \theta_t - \epsilon\, P(\theta_t)\, \nabla f(\theta_t),$$

  where $P(\theta)$ is positive definite.
- "Good" P approximates curvature (Newton-like) while remaining cheap and stable.
- Information geometry suggests $P(\theta) \approx F(\theta)^{-1}$ (Fisher) when the objective is likelihood / KL-based.

NATURAL GRADIENT AS A PRINCIPLED PRECONDITIONER

- For negative log-likelihood objectives, the Fisher matrix often behaves like a "statistical Hessian".
- Natural gradient:

$$\theta_{t+1} = \theta_t - \epsilon\, F(\theta_t)^{-1} \nabla f(\theta_t)$$

  is Newton-like, but invariant under reparametrization.
- It is also the first-order solution of a KL trust-region step (Part 4).
- Practical view: "use curvature measured in distribution space, not parameter space".

Rule of thumb: if the training objective compares distributions, Fisher-based geometry is a natural candidate.

APPROXIMATIONS IN PRACTICE (DEEP LEARNING)

- Full Fisher/Hessian is too large: we use structured approximations.
- Common choices: diagonal / block-diagonal / low-rank / Kronecker-factored approximations.
- Many “adaptive” optimizers can be interpreted as diagonal or blockwise preconditioners (often heuristic).
- The IG perspective helps distinguish: *which geometry are we approximating?* (Euclidean vs Fisher vs other).

CURVATURE: FISHER, HESSIAN, GAUSS–NEWTON (CONCEPTUAL MAP)

- For log-likelihood objectives, the Hessian splits into “data-fit curvature” + “model curvature” terms.
- Fisher is always positive semidefinite (a covariance), hence stable as a curvature surrogate.
- In many settings, Fisher aligns with a generalized Gauss–Newton approximation.
- Practical takeaway: Fisher-type preconditioning often avoids the instability of indefinite Hessians.

CHOOSING THE GEOMETRY TO MATCH THE LOSS

- If your loss is KL / cross-entropy-like, Fisher geometry is locally canonical (2nd-order term of KL).
- If your loss is a Bregman divergence, mirror descent is the natural first-order method.
- If your loss is not KL-like (e.g. transport distances), a different geometry may be more faithful.
- Main message: the loss defines the notion of "small change" — and that defines your optimizer.

WHAT DO DUAL CONNECTIONS MEAN IN ML TERMS?

- Dual connections encode two compatible affine structures on the same model space.
- In exponential families: e-affine = natural parameters θ (linear structure in $\log p$); m-affine = expectation parameters η (linear structure under mixing).
- Algorithmically: some updates are simpler in the "primal" chart (natural params), others in the "dual" chart (expectations / moments).
- The dual pair (∇, ∇^*) is the geometric reason why "update in one space, map back" (mirror descent / VI) works so cleanly.

Unifying view: dual connections formalize the two-sided nature of learning: fit parameters ↔ match moments.

A PRACTICAL WORKFLOW (HOW TO USE IG IDEAS)

- Step 1: identify the divergence / loss that measures model discrepancy in your task.
- Step 2: take its local quadratic term → metric g (often Fisher if KL-based).
- Step 3: decide if a dual flat structure exists (exponential-family-like) → use (θ, η) and Bregman tools.
- Step 4: pick an optimizer consistent with that geometry (NG, MD, trust regions, ...).

PART 5 RECAP

- Fisher/natural gradient provides a principled curvature and invariance, useful for preconditioning and quasi-2nd-order ideas.
- “Right geometry” depends on the chosen loss; divergences induce metrics + dual connections.
- Dual connections explain why many algorithms have a primal/dual form (parameters vs moments, θ vs η).

End of the core story.

BIG PICTURE: LOSS ⟶ GEOMETRY ⟶ ALGORITHMS

Start from what you measure

Loss / Divergence

KL / cross-entropy
Bregman losses
other discrepancies

$D(p||q)$

⟶

Induces local & affine structure

Geometry

metric g (2nd order)
dual connections ∇ / ∇^* (3rd order)
(dually flat ⇒ θ / η)

Legendre duality
Bregman geometry

⟶

Choose updates consistent with geometry

Algorithms

Natural gradient (Fisher)
Mirror descent (Bregman)
Trust regions (KL)
Preconditioning / 2nd-order approximations

Takeaway: pick the geometry that matches your notion of discrepancy — dual connections explain the "primal/dual" structure behind many learning & inference updates.