The Information Geometry of Optimization

Gianluca Covini

Presentation for the exam of *Optimization*

References

- **Reference for the main result:** G. Raskutti, S. Mukherjee, *The Information Geometry of Mirror Descent*, IEEE Transactions on Information Theory, 2015
- **Reference for general explanation of information geometry:** F. Nielsen, *An Elementary Introduction to Information Geometry*, Entropy, 2020

Motivations

Setting

Setting

$$L : \Theta \to \mathbb{R}; \quad \text{find} \quad \theta^* = \arg\min_{\theta \in \Theta} L(\theta)$$

Setting

Setting

$$L : \Theta \to \mathbb{R}; \quad \text{find} \quad \theta^* = \arg\min_{\theta \in \Theta} L(\theta)$$

Possible Approach

$$\text{Gradient Descent (GD):} \quad \theta_{t+1} = \theta_t - \alpha_t \nabla L(\theta_t)$$

Setting

$$L : \Theta \to \mathbb{R}; \quad \text{find} \quad \theta^* = \arg\min_{\theta \in \Theta} L(\theta)$$

Possible Approach

$$\text{Gradient Descent (GD):} \quad \theta_{t+1} = \theta_t - \alpha_t \nabla L(\theta_t)$$

Problem

GD is **coordinate-dependent**

$$\theta_{t+1} = \theta_t - \alpha_t \nabla L_\theta(\theta_t) \quad \xrightarrow[L_\eta(\eta)=L_\theta(\theta(\eta))]{\theta=\theta(\eta)} \quad \eta_{t+1} = \eta_t - \alpha_t \nabla L_\eta(\eta_t)$$

→ In general, $\{\theta_t\}_t \neq \{\eta_t\}_t$.

Setting

Setting

$$L : \Theta \to \mathbb{R}; \quad \text{find} \quad \theta^* = \arg\min_{\theta \in \Theta} L(\theta)$$

Possible Approach

$$\text{Gradient Descent (GD):} \quad \theta_{t+1} = \theta_t - \alpha_t \nabla L(\theta_t)$$

Problem

GD is **coordinate-dependent**

$$\theta_{t+1} = \theta_t - \alpha_t \nabla L_\theta(\theta_t) \quad \xrightarrow[L_\eta(\eta) = L_\theta(\theta(\eta))]{\theta = \theta(\eta)} \quad \eta_{t+1} = \eta_t - \alpha_t \nabla L_\eta(\eta_t)$$

→ In general, $\{\theta_t\}_t \neq \{\eta_t\}_t$.

→ GD does not take into account the **geometry** of the problem.

Goal

Goal of the Presentation

Show a “geometry-wise” approach to optimization for a wide class of spaces of applicative interest.

Goal

Goal of the Presentation

Show a "geometry-wise" approach to optimization for a wide class of spaces of applicative interest.

Which space?

find $\quad p^* = \arg\min_{p \in M} L(p)$ when M is an **information manifold**.

Goal

Goal of the Presentation

Show a “geometry-wise” approach to optimization for a wide class of spaces of applicative interest.

Which space?

find $\quad p^* = \arg\min_{p \in M} L(p)$ when M is an **information manifold**.

$\rightarrow$ Riemannian manifolds with additional structure.

$\rightarrow$ Aim to geometrically represent how information passes from data to models.

Motivating Example

Example

$$M := \{p_\theta = \mathcal{N}(\theta) \mid \theta = \left(\Sigma^{-1}\mu, -\Sigma^{-1}/2\right) \in \Theta\}$$

with appropriate manifold structure is an **information manifold**.

Motivating Example

Example

$$M := \{p_\theta = \mathcal{N}(\theta) \mid \theta = \left(\Sigma^{-1}\mu, -\Sigma^{-1}/2\right) \in \Theta\}$$

with appropriate manifold structure is an **information manifold**.

Optimization Problem

Given x data, we may want to find θ^* the **maximum likelihood estimator**

$$\begin{aligned} &\text{find} \quad \theta^* \\ &\text{solving} \quad \min_{\theta\in\Theta} \underbrace{\ell(\theta; x)}_{\substack{\text{negative}\\ \text{log-likelihood}}} = \min_{p_\theta \in M} -\log(p_\theta(x)) \end{aligned}$$

Table of contents

Information Geometry

Manifolds

Manifold

A D**-dimensional manifold** is a topological space (locally) homeomorphic to an open set of $\mathbb{R}^D$.

$\rightarrow$ We will consider global homeomorphism.

We can define a set of coordinates:

$$\begin{aligned} \theta : & M \rightarrow \Theta \subseteq \mathbb{R}^D \\ & p \mapsto \theta(p) = (\theta_1(p), \ldots, \theta_D(p)) \end{aligned}$$

where θ is a homeomorphism (continuous bijection). >>

Differential Manifold

We can define functions $L : M \to \mathbb{R}$ and $\frac{\partial L}{\partial \theta_i}(p) := \frac{\partial (L \circ \theta^{-1})}{\partial x_i}(\theta(p))$

Tangent Space

$\forall p \in M$, we associate a **tangent space** T_pM, which can be seen as the space of directional derivatives

$$T_pM := \{v : C^\infty(M) \to \mathbb{R} \mid v \text{ linear; } v(fg) = v(f)g + fv(g)\}$$

→ T_pM is a D-dimensional vector space with natural basis corresponding to partial derivatives $\mathcal{B} := \{e_i, \quad i = 1, \dots, D\}$.

→ We can define a **vector field** as a function $X : p \in M \mapsto v \in T_pM$

>>

Riemannian Manifolds

We want to define an **inner product** g_p on T_pM.

Riemannian Manifolds

We want to define an **inner product** g_p on T_pM.

Two examples of information manifolds we will see:

Statistical Manifolds

In general, we can consider the manifolds of parametric families induced by the **Fisher information** $M := \{p_\theta \mid \theta \in \Theta\}, \quad \mathcal{I}(\theta) = \left(\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\ell(\theta;x)\frac{\partial}{\partial\theta_j}\ell(\theta;x)\right]\right)$

$\rightarrow \quad g_p(u,v) = u^T\mathcal{I}(\theta(p))v \quad \forall u,v \in T_pM$

Riemannian Manifolds

We want to define an **inner product** g_p on T_pM.

Two examples of information manifolds we will see:

Statistical Manifolds

In general, we can consider the manifolds of parametric families induced by the **Fisher information** $M := \{p_\theta \mid \theta \in \Theta\}, \quad \mathcal{I}(\theta) = \left(\mathbb{E}\left[\frac{\partial}{\partial \theta_i}\ell(\theta; x)\frac{\partial}{\partial \theta_j}\ell(\theta; x)\right]\right)$

$\rightarrow \quad g_p(u, v) = u^T \mathcal{I}(\theta(p)) v \quad \forall u, v \in T_pM$

Bregman Manifolds

The inner product can also be induced by **Bregman divergence**.

$F : \Theta \rightarrow \mathbb{R}$ *mirror map* $\rightsquigarrow \quad B_F(\theta \mid \theta') := F(\theta) - F(\theta') - (\theta - \theta')^T \nabla F(\theta')$

$\rightarrow \quad g_p(u, v) = u^T \nabla^2 F(\theta(p)) v \quad \forall u, v \in T_pM$

$\rightarrow$ Equivalent for exponential parametric families (e.g. Gaussians).

Riemannian Manifolds

We want to define an **inner product** g_p on T_pM.

Statistical Manifolds

In general, we can consider the manifolds of parametric families induced by the **Fisher information** $M := \{p_\theta \mid \theta \in \Theta\}, \quad \mathcal{I}(\theta) = \left(\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\ell(\theta;x)\frac{\partial}{\partial\theta_j}\ell(\theta;x)\right]\right)$

$$\rightarrow \quad g_p(u,v) = u^T\mathcal{I}(\theta(p))v \quad \forall u,v \in T_pM$$

Bregman Manifolds

The inner product can also be induced by **Bregman divergence**.

$$F : \Theta \rightarrow \mathbb{R} \quad \textit{mirror map} \quad \rightsquigarrow \quad B_F(\theta \mid \theta') := F(\theta) - F(\theta') - (\theta - \theta')^T\nabla F(\theta')$$

$$\rightarrow \quad g_p(u,v) = u^T\nabla^2 F(\theta(p))v \quad \forall u,v \in T_pM$$

Riemannian Manifolds

(M,g) as defined is a Riemannian manifold.

Note: from now on, we will consider information manifolds as induced by Bregman divergence.

To define an information manifold we also need more structure

Affine Connection

We define the **affine connection** a $\nabla : (X, Y) \mapsto \nabla_X Y$ vector field.

$\rightarrow$ For Bregman manifolds, $\nabla_{e_i} e_j = 0$ for every $e_i, e_j \in \mathcal{B}$ natural basis of $T_p M$.

Note: from now on, we will consider information manifolds as induced by Bregman divergence.

To define an information manifold we also need more structure

Affine Connection

We define the **affine connection** a $\nabla : (X, Y) \mapsto \nabla_X Y$ vector field.

$\rightarrow$ For Bregman manifolds, $\nabla_{e_i} e_j = 0$ for every $e_i, e_j \in \mathcal{B}$ natural basis of $T_p M$.

We also need a **dual structure**,

$$\nabla^* : (X, Y) \mapsto \nabla^*_X Y$$

Dual

Three ways to see duality

$$\nabla \longrightarrow \nabla^* \quad \text{s.t.} \quad X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle$$

$$\nabla \text{ induced by } B_F \longrightarrow \nabla^* \text{ induced by } B^*_F(\theta \mid \theta') := B_F(\theta' \mid \theta)$$

$$\nabla \text{ induced by } B_F \longrightarrow \nabla^* \text{ induced by } B_{F^*}$$

$$\text{where } F^*(\eta) := \sup_{\theta \in \Theta}\{\theta^T \eta - F(\theta)\}$$

Dual

Three ways to see duality

$$\begin{aligned} \nabla \quad &\longrightarrow \nabla^* \quad \text{s.t.} \quad X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle \\ \nabla \text{ induced by } B_F &\longrightarrow \nabla^* \text{ induced by } B^*_F(\theta \mid \theta') := B_F(\theta' \mid \theta) \\ \nabla \text{ induced by } B_F &\longrightarrow \nabla^* \text{ induced by } B_{F^*} \\ &\qquad \text{where } F^*(\eta) := \sup_{\theta \in \Theta}\{\theta^T \eta - F(\theta)\} \end{aligned}$$

Information manifold

$(M, g_P, \nabla, \nabla^*)$ induced by B_F is an **information manifold**.

Dual

Three ways to see duality

$$\begin{aligned} \nabla \quad &\longrightarrow \nabla^* \quad \text{s.t.} \quad X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle \\ \nabla \text{ induced by } B_F &\longrightarrow \nabla^* \text{ induced by } B^*_F(\theta \mid \theta') := B_F(\theta' \mid \theta) \\ \nabla \text{ induced by } B_F &\longrightarrow \nabla^* \text{ induced by } B_{F^*} \\ &\qquad \text{where } F^*(\eta) := \sup_{\theta \in \Theta} \{\theta^T \eta - F(\theta)\} \end{aligned}$$

Information manifold

$(M, g_P, \nabla, \nabla^*)$ induced by B_F is an **information manifold**.

What does duality imply? It defines two sets of **coordinates**

$$\theta = \nabla F^*(\eta) \longleftrightarrow \eta = \nabla F(\theta)$$

Optimization on Information Manifolds

Riemann Gradient Descent

Riemannian Gradient Descent

Given (M, g) Riemannian manifold.

- → **Generalization of GD** in the sense of optimizing by moving in the direction of steepest descent;
- → We define the map $\exp_p(v)$ which gives the arrival point after a unit of time of the shortest curve starting from p with velocity v.

$$\text{RGD:} \quad p_{t+1} = \exp_{p_t}(-\alpha_t \nabla_M L(p_t))$$

>>

NGD and MD

Problem

$\exp_p(v)$ is computationally intractable.

NGD and MD

Problem

$\exp_p(v)$ is computationally intractable.

On the **Bregman manifold** (M, F),

Natural Gradient Descent (NGD)

We can replace $\exp_p(v)$ with its first-order Taylor approximation $\exp_p(v) \approx p + v$.

$$\text{NGD:} \quad \theta_{t+1} = \theta_t - \alpha_t \underbrace{(\nabla^2_\theta F(\theta_t))^{-1} \nabla_\theta (L_\theta(\theta_t))}_{\nabla^{(NG)} L_\theta(\theta_t) \text{ natural gradient}}$$

NGD and MD

Problem

$\exp_p(v)$ is computationally intractable.

On the **Bregman manifold** (M, F),

Natural Gradient Descent (NGD)

We can replace $\exp_p(v)$ with its first-order Taylor approximation $\exp_p(v) \approx p + v$.

$$\text{NGD:} \quad \theta_{t+1} = \theta_t - \alpha_t \underbrace{(\nabla_\theta^2 F(\theta_t))^{-1} \nabla_\theta (L_\theta(\theta_t))}_{\nabla^{(NG)} L_\theta(\theta_t) \text{ natural gradient}}$$

Mirror Descent (MD)

$$\text{MD:} \quad \theta_{t+1} = \underset{\theta \in \Theta}{\arg\min} \left\{ \theta^T \nabla_\theta L_\theta(\theta_t) + \frac{1}{\alpha_t} B_F(\theta \mid \theta_t) \right\}$$

Equivalence Result

Theorem [Raskutti, Mukherjee]

Given an information manifold (M, g, ∇, ∇^*) induced by a Bregman divergence B_F, **MD on** (M, F) **is equivalent to NGD in the dual space** (M, F^*).

Equivalence Result

Theorem [Raskutti, Mukherjee]

Given an information manifold (M, g, ∇, ∇^*) induced by a Bregman divergence B_F, **MD on** (M, F) **is equivalent to NGD in the dual space** (M, F^*).

Proof

$$\text{MD:} \quad \theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \theta^T \nabla L_\theta(\theta_t) + \frac{1}{\alpha_t} B_F(\theta \mid \theta_t) \right\}$$

Equivalence Result

Theorem [Raskutti, Mukherjee]

Given an information manifold (M, g, ∇, ∇^*) induced by a Bregman divergence B_F, **MD on** (M, F) **is equivalent to NGD in the dual space** (M, F^*).

Proof

$$\text{MD:} \quad \theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \theta^T \nabla L_\theta(\theta_t) + \frac{1}{\alpha_t} B_F(\theta \mid \theta_t) \right\}$$

Finding the minimum by differentiation yields the step:

$$\nabla F(\theta_{t+1}) = \nabla F(\theta_t) - \alpha_t \nabla_\theta L(\theta_t)$$

Equivalence Result

Theorem [Raskutti, Mukherjee]

Given an information manifold (M, g, ∇, ∇^*) induced by a Bregman divergence B_F, **MD on** (M, F) **is equivalent to NGD in the dual space** (M, F^*).

Proof

$$\text{MD:} \quad \theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \theta^T \nabla L_\theta(\theta_t) + \frac{1}{\alpha_t} B_F(\theta \mid \theta_t) \right\}$$

Finding the minimum by differentiation yields the step:

$$\nabla F(\theta_{t+1}) = \nabla F(\theta_t) - \alpha_t \nabla_\theta L(\theta_t)$$

Dual change of variable: $\eta = \nabla F(\theta), \quad \theta = \nabla F^*(\eta),$

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\theta L(\nabla F^*(\eta_t))$$

Equivalence Result

Proof

$$\text{MD:} \quad \theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \theta^T \nabla L_\theta(\theta_t) + \frac{1}{\alpha_t} B_F(\theta \mid \theta_t) \right\}$$

Finding the minimum by differentiation yields the step:

$$\nabla F(\theta_{t+1}) = \nabla F(\theta_t) - \alpha_t \nabla_\theta L(\theta_t)$$

Dual change of variable: $\eta = \nabla F(\theta), \quad \theta = \nabla F^*(\eta)$,

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\theta L(\nabla F^*(\eta_t))$$

Chain rule: $\nabla_\eta L(\nabla F^*(\eta)) = \nabla_\eta^2 F^*(\eta) \nabla_\theta L(\nabla F^*(\eta))$

Equivalence Result

Proof

$$\text{MD:}\quad \theta_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \theta^T \nabla L_\theta(\theta_t) + \frac{1}{\alpha_t} B_F(\theta \mid \theta_t) \right\}$$

Finding the minimum by differentiation yields the step:

$$\nabla F(\theta_{t+1}) = \nabla F(\theta_t) - \alpha_t \nabla_\theta L(\theta_t)$$

Dual change of variable: $\eta = \nabla F(\theta), \quad \theta = \nabla F^*(\eta)$,

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\theta L(\nabla F^*(\eta_t))$$

Chain rule: $\nabla_\eta L(\nabla F^*(\eta)) = \nabla_\eta^2 F^*(\eta) \nabla_\theta L(\nabla F^*(\eta))$

Therefore, $\qquad \eta_{t+1} = \eta_t - \alpha_t (\nabla^2 F^*(\eta_t))^{-1} \nabla_\eta L(\nabla F^*(\eta_t))$

which corresponds to the natural gradient descent step. □

Application

Back to our motivating example,

$$M := \{p_\theta = \mathcal{N}(\theta) \mid \theta = \left(\Sigma^{-1}\mu, -\Sigma^{-1}/2\right) \in \Theta\} \text{ with } F(\theta) := \frac{1}{2}\|\theta\|_2^2$$

We want to find the MLE θ^* given x data:

$$\text{find} \quad \theta^* = \arg\min_{\theta\in\Theta} \ell(\theta; x) = \arg\min_{p_\theta \in M} -\log(p_\theta(x))$$

Application

- NGD moves in the direction of steepest descent of ℓ and asymptotically achieves the minimum possible asymptotic variance (CR bound) **but** it requires $\nabla^2 F$;
- for MD we don't have guarantees of moving in direction of steepest descent and of achieving asymptotical CR bound, **but** it is a I-order method.

$\rightarrow$ The equivalence result guarantees that we have a I-order method that achieves CR bound.

Conclusion

Conclusion

Takeaways

$\rightarrow$ Optimization is strongly influenced by space geometry;

Conclusion

Takeaways

→ Optimization is strongly influenced by space geometry;

→ we saw information manifolds induced by parametric families and by Bregman divergence;

Conclusion

Takeaways

- → Optimization is strongly influenced by space geometry;
- → we saw information manifolds induced by parametric families and by Bregman divergence;
- → on information manifolds optimization can be performed through NGD and MD;

Conclusion

Takeaways

- → Optimization is strongly influenced by space geometry;
- → we saw information manifolds induced by parametric families and by Bregman divergence;
- → on information manifolds optimization can be performed through NGD and MD;
- → the two are equivalent on Bregman manifolds.

Conclusion

Takeaways

- → Optimization is strongly influenced by space geometry;
- → we saw information manifolds induced by parametric families and by Bregman divergence;
- → on information manifolds optimization can be performed through NGD and MD;
- → the two are equivalent on Bregman manifolds.

Thank you for your attention!