Plan $\begin{cases} \text{Optimality Condition on a manifold} \\ \text{Convergence of Riemannian GD} \end{cases}$

We discuss optimality using an example:

$$① \begin{cases} \min\limits_{X \in \mathbb{R}^{n\times p}} f(X) = \operatorname{tr}(X^T A X) & A^T = A \in \mathbb{R}^{n\times n} \\ \text{s.t.} \quad X^T X = I \end{cases}$$

$$\Longleftrightarrow \quad ② \quad \min_{X \in M} f(X), \quad M = St(n,p) = \{X \in \mathbb{R}^{n\times p} : X^T X = I\}$$

I. Necessary but not sufficient optimality condition for ①

can be derived from saddle point of Lagrangian

$$L(X, \Lambda) = f(X) - \langle \Lambda, X^T X - I \rangle$$

$X \in \mathbb{R}^{n\times p}$ $\quad \Lambda \in \mathbb{R}^{p\times p}$ $\quad \langle U, V \rangle = \sum_i \sum_j U_{ij} V_{ij}$

Regard $f(X)$ as a function on $\mathbb{R}^{n\times p}$

$$\frac{\partial L}{\partial X} = \frac{\partial}{\partial X} f(X) - \frac{\partial}{\partial X} \langle \Lambda, X^T X - I \rangle$$

$$= \frac{\partial}{\partial X} \langle X, AX \rangle - \frac{\partial}{\partial X} \langle \Lambda, X^T X \rangle$$

$$= 2AX - (X\Lambda + X\Lambda^T)$$

How to calculate $\frac{\partial}{\partial X} \langle \Lambda, X^T X \rangle$: $\quad \operatorname{tr}(A^T B) = \langle A, B \rangle$

1) $\langle \Lambda, Y^T X \rangle = \operatorname{tr}(\Lambda^T Y^T X) = \operatorname{tr}((Y\Lambda)^T X) = \langle Y\Lambda, X \rangle$

$$\Rightarrow \frac{\partial}{\partial X} \langle \Lambda, Y^T X \rangle = Y\Lambda$$

2) $\langle \Lambda, Y^T X\rangle = tr(X^T Y \Lambda) \underset{\downarrow}{=} tr(\Lambda X^T Y) = \langle X\Lambda^T, Y\rangle$

$tr(ABC) = tr(CAB)$

$\Rightarrow \frac{\partial}{\partial Y}\langle \Lambda, Y^T X\rangle = X\Lambda^T$

3) $\frac{\partial}{\partial X}\langle \Lambda, X^T X\rangle = \frac{\partial}{\partial X}\langle \Lambda, Y^T X\rangle\Big|_{Y=X} + \frac{\partial}{\partial Y}\langle \Lambda, Y^T X\rangle\Big|_{Y=X}$

$= X\Lambda + X\Lambda^T$

$L(X, \Lambda) = f(X) - \langle \Lambda, X^T X - I\rangle$

Saddle Point $\begin{cases} \frac{\partial L}{\partial X} = \frac{\partial}{\partial X} f(X) - (X\Lambda + X\Lambda^T) = 0 \\ \frac{\partial L}{\partial \Lambda} = X^T X - I = 0 \end{cases}$

II. Necessary but not sufficient optimality condition for ②

- a manifold $\underset{St(n,p)}{\underset{\parallel}{M}} \subset \underset{\mathbb{R}^{n\times p}}{\underset{\parallel}{\mathcal{E}}}$

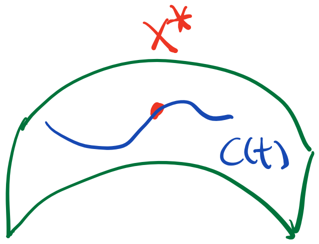
- a curve $c: (-\varepsilon, \varepsilon) \to M$
  $t \mapsto c(t)$

To understand/calculate $c'(t)$, we identify it with a natural extension

$c: \mathbb{R} \to \mathcal{E}$

$t \mapsto c(t)$

- If $x_* \in M$ is the minimizer of $f(x)$ on M, then consider any curve $c(t)$ on M with $c(0) = x_*$.



$f \circ c(t)$ has a minimizer at $t=0$

$$\Rightarrow \frac{d}{dt}[f \circ c]\Big|_{t=0} = 0$$

- Recall the definition of differential of $f: M \to \mathbb{R}$ is a linear map

$$Df(x): T_x M \to \mathbb{R}$$
$$v \mapsto Df(x)[v] \overset{\text{def}}{=} \frac{d}{dt} f(\gamma(t))\Big|_{t=0}$$

where $\gamma(t)$ is any curve on M with $\begin{cases} \gamma(0) = x \\ \gamma'(0) = v \end{cases}$

- Theorem (Necessary Optimality Condition on M)

(a) $\begin{cases} \frac{d}{dt}[f \circ c]\Big|_{t=0} = 0 \\ \text{for any curve } c(t) \text{ on } M \text{ with } c'(0) = x_* \end{cases}$

$\Leftrightarrow$ (b) $Df(x_*)[v] = 0, \ \forall v \in T_{x_*} M$

$\Leftrightarrow$ (c) $\operatorname{grad} f(x_*) = 0$

Proof: (b) $\Leftrightarrow$ (c)

def of Riemannian Grad

$$g_{x_*}(\operatorname{grad} f(x_*), v) = Df(x_*)[v], \ \forall v \in T_{x_*} M$$

Since $\mathrm{grad} f(X_*) \in T_* M$, we pick $v = \mathrm{grad} f(X_*)$,

then $g_{X_*}(\mathrm{grad} f(X_*), \mathrm{grad} f(X_*)) = 0$

$\Rightarrow \mathrm{grad} f(X_*) = 0$

positive-definiteness of metric

(a) $\Leftrightarrow$ (b)

(a) $\left\{ \begin{array}{l} \frac{d}{dt}[f \circ c]\Big|_{t=0} = 0 \\ \text{for any curve } c(t) \text{ on } M \text{ with } c(0) = X_* \end{array} \right.$

The general definition of tangent space of a manifold $M \subseteq \mathcal{E}$

$$T_x M = \{ c'(0) \mid c: (-\varepsilon, \varepsilon) \to M \text{ is smooth } \& \ c(0) = x \}$$

For any curve $\gamma(t)$ with $\gamma(0) = X_*$, let $V = \gamma'(0)$

then $Df(x)[V] \overset{def}{=} \frac{d}{dt} f(\gamma(t))\Big|_{t=0} = 0$

arbitrariness of such a curve $\Leftrightarrow \forall V \in T_{X_*} M$

- So the optimality on M is $\mathrm{grad} f(X_*) = 0$

$$g_{X_*}(\mathrm{grad} f(X_*), V) = Df(X_*)[V] = D\bar{f}(X_*)[V]$$
$$= \langle \nabla \bar{f}(X_*), V \rangle$$

$f: M \to \mathbb{R}$ has an extension $\bar{f}: \mathbb{R}^{n \times p} \to \mathbb{R}$

$X \mapsto \mathrm{tr}(X^T A X)$ $\qquad$ $X \mapsto \mathrm{tr}(X^T A X)$

In particular, if we choose $g_x(\,,\,)$ to be $\langle\,,\,\rangle$

then $\operatorname{grad} f(X_*)$ is Proj of $\nabla \bar{f}(X_*)$

on $T_{X_*}M$

Previously, we have computed

$$\forall\, Y \in \mathbb{R}^{n\times p},\quad P_{T_XM}(Y) = (I - XX^T)Y + \operatorname{skew}(X^TY)$$

$$\boxed{T_X St(n,p) = \{ V \in \mathbb{R}^{n\times p} : X^TV + V^TX = 0 \} \subseteq \mathbb{R}^{n\times p}}$$

So
$$\begin{aligned}
\operatorname{grad} f(X) &= (I - XX^T)\nabla \bar{f}(X) + \frac{X^T\nabla \bar{f}(X) - \nabla \bar{f}(X)^TX}{2} \\
&= (I - XX^T)2AX + X^TAX - X^TAX \\
&= (I - XX^T)2AX \quad ②
\end{aligned}$$

Let's compare it to

Optimality via $L(X,\Lambda) = f(X) - \langle \Lambda, X^TX - I\rangle$

① Saddle Point
$$\begin{cases} \dfrac{\partial L}{\partial X} = \dfrac{\partial}{\partial X} f(X) - (X\Lambda + X\Lambda^T) = 0 \\[2mm] \dfrac{\partial L}{\partial \Lambda} = X^TX - I = 0 \end{cases}$$

Claim : ① $\iff$ ② $\left(\operatorname{grad} f(X) = P_{T_XM}(\nabla \bar{f}(X)) = 0\right)$

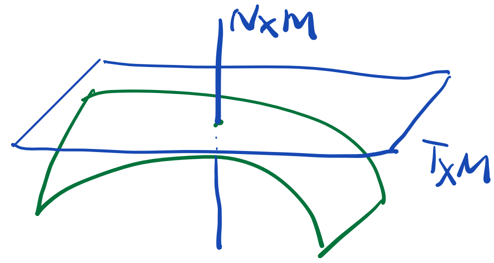
Proof : Previously we derived the normal space

$$N_X St(n,p) = \{ XS : S^T = S,\ S \in \mathbb{R}^{p\times p} \}$$

T_XM is a subspace in $\mathcal{E}$

N_XM is the orthogonal complement of T_XM

and $\mathcal{E} = T_xM \oplus N_xM$



$$\frac{\partial}{\partial x} f(x) - (X\Lambda + X\Lambda^T) = 0$$

$$\Leftrightarrow \frac{\partial}{\partial x} f(x) = X(\underbrace{\Lambda + \Lambda^T}_{S}) \in N_x St(n,p)$$

$\Leftrightarrow$ Proj of $\frac{\partial}{\partial x} f(x)$ onto T_xM is 0

$\Leftrightarrow$ ② because $\frac{\partial}{\partial x} f(x)$ in ① is $\nabla \bar{f}(x)$ in ②

- Remark & Observation :

1) When using Euclidean metric $g_x(\,,\,) = \langle\,,\,\rangle$

$\operatorname{grad} f(x) = P_{T_xM}\left(\frac{\partial}{\partial x} f(x)\right)$, so ① $\Leftrightarrow$ ②

2) If using a generic metric g, then

$\operatorname{grad} f(x) \neq P_{T_xM}\left(\frac{\partial}{\partial x} f(x)\right)$.

3) Intuitively, optimality conditions ① & ②

should be equivalent, regardless of metric.

But if $\operatorname{grad} f(x) \neq P_{T_xM}\left(\frac{\partial}{\partial x} f(x)\right)$,

why ① $\Leftrightarrow$ ② ?

Answer: ① $\Leftrightarrow$ $P_{T_xM}\left(\frac{\partial}{\partial x} f(x)\right) = 0 \Leftrightarrow \frac{\partial}{\partial x} f(x) \in N_xM$

② $\Leftrightarrow$ $\operatorname{grad} f(x) = 0$

① $\Rightarrow g_x(\text{grad} f(x), v) = \langle \nabla \bar{f}(x), v \rangle = 0, \forall v \in T_x M$

$\Rightarrow g_x(\text{grad} f(x), v) = 0$

$\Rightarrow g_x(\text{grad} f(x), \text{grad} f(x)) = 0 \Rightarrow \text{grad} f(x) = 0$

② $\Rightarrow \forall v \in T_x M, \; g_x(\text{grad} f(x), v) = 0$

$\Rightarrow \langle \nabla \bar{f}(x), v \rangle = g_x(\text{grad} f(x), v) = 0$

$\Rightarrow \nabla \bar{f}(x) \in N_x M$

$\Rightarrow$ Proj of $\nabla \bar{f}(x)$ onto $T_x M$ is 0

Riemannian Gradient Descent for $\min_{x \in M} f(x)$

$$x_{k+1} = R_{x_k}(-\eta_k \, \text{grad} f(x_k))$$

We expect x_k converge to x_* s.t. $\text{grad} f(x_*) = 0$

just a critical point
not necessarily a minimizer

Step size rule/method:

1) a constant step size $\eta_k = \eta$

2) "optimal" step size

$$\eta_k = \underset{t}{\text{argmin}} \; h(t) = R_{x_k}(-t \, \text{grad} f(x_k))$$

3) Line search by backtracking

Start with step size t_0, iteratively reduce it to $t_i = \rho t_{i-1}$ for some $\rho \in (0,1)$ until some conditions are satisfied.

Ideas/Steps to show the convergence:

① Assume $f(x) \geq D$, $\forall x \in M$

② Show sufficient decrease

$$f(x_k) - f(x_{k+1}) \geq c \, \|\mathrm{grad}\, f(x_k)\|^2$$

only possible with certain step sizes

assumptions and f & (M, g)

③ Then we can show $\lim_{k \to \infty} \|\mathrm{grad}\, f(x_k)\| = 0$

$$0 \leq \sum_{k=0}^{N} [f(x_k) - f(x_{k+1})] = f(x_0) - f(x_N) \leq f(x_0) - D$$

$\Rightarrow \sum_{k=0}^{\infty} [f(x_k) - f(x_{k+1})]$ is finite

$\Rightarrow c \, \|\mathrm{grad}\, f(x_k)\|^2 \leq f(x_k) - f(x_{k+1}) \to 0, \quad k \to \infty$

$$f(x_0) - D \geq f(x_0) - f(x_N)$$

$$= \sum_{k=0}^{N} \left[f(x_k) - f(x_{k+1}) \right]$$

$$\geq N \, c \min_{0<k\leq N} \| \mathrm{grad} f(x_k) \|^2$$

$$\Rightarrow \min_{0<k\leq N} \| \mathrm{grad} f(x_N) \| \leq \sqrt{\frac{f(x_0) - D}{c}} \frac{1}{\sqrt{N}}$$

④ How to get sufficient decrease

$$f(x_k) - f(x_{k+1}) \geq c \, \| \mathrm{grad} f(x_k) \|^2$$

1) For special f, R_x,

$$v \in T_x M, \quad f(R_x(v)) \leq f(x) + \langle \mathrm{grad} f(x), v \rangle + \frac{L}{2} \|v\|^2$$

2) With constant step size $\eta_k = \frac{1}{L}$, can show

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \| \mathrm{grad} f(x_k) \|^2$$

Proof: $v = -\frac{1}{L} \mathrm{grad} f(x_k)$