

# 量子 Fisher 信息矩阵 vs 经典 Fisher 信息矩阵

从 KL 散度到 Fubini-Study 距离, 从自然梯度到随机重构的自洽推导

## 摘要

经典信息几何中, KL 散度在小参数扰动下的二阶展开诱导出 Fisher-Rao 度量, 其度量张量即 Fisher 信息矩阵 (FIM); 对应的“最速下降”给出自然梯度法。量子纯态的物理态是射线 (ray), 存全局相位与整体归一化的规范冗余, 因而不能直接把“对输出的欧氏内积/协方差”当作度量; 正确的距离应定义在射影 Hilbert 空间上, 即 Fubini-Study (FS) 距离。FS 距离的二阶展开给出量子 Fisher 信息矩阵 (又称量子几何张量, QGT): 其实部是 FS 黎曼度量, 虚部是 Berry 曲率。进一步地, 在以 FS 范数约束的最速下降问题中, 参数更新满足线性方程  $S\delta\theta = -\eta g$ , 这正是变分蒙特卡洛与神经网络量子态训练中常用的随机重构 (SR) / 量子自然梯度 (QNG) 更新。本文从经典部分开始, 给出 KL 一阶项为零、二阶项等于 score 协方差的完整证明, 并在清晰区分“概率输出”和“复振幅输出”的基础上, 逐步引入 FS 距离、量子 Fisher 信息矩阵的规范不变定义及其与 SR 的几何推导。文末用两张对照表总结经典与量子概念的对应关系, 以及常见优化算法与其隐含的度量/流形解释 (包括 Adam/RMSProp 等对角近似)。

## 目录

<b>1 经典概率分布</b>	<b>1</b>
1.1 从 KL 散度到 Fisher 信息矩阵 . . . . .	1
1.2 Score 向量 . . . . .	3
1.3 流形上的最速下降: 自然梯度 . . . . .	3
<b>2 从经典概率分布到量子态: 量子 Fisher 信息矩阵</b>	<b>3</b>
2.1 从概率模型类比到量子态 . . . . .	3
2.2 从 Fubini-Study 距离到量子 Fisher 信息矩阵 . . . . .	4
2.3 流形上的最速下降: 随机重构 / 量子自然梯度 . . . . .	5
<b>A 优化算法的度量视角</b>	<b>5</b>

## 1 经典概率分布

### 1.1 从 KL 散度到 Fisher 信息矩阵

设  $p_{\boldsymbol{\theta}}(x)$  是由一组参数  $\boldsymbol{\theta} = (\theta^1, \dots, \theta^d)$  参数化的概率分布 (离散与连续情形统一记作  $\sum_x$ )。两分布之间的差异由 Kullback-Leibler (KL) 散度刻画:

$$D_{\text{KL}}(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'}) := \sum_x p_{\boldsymbol{\theta}}(x) \ln \frac{p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}'}(x)} = \sum_x p_{\boldsymbol{\theta}}(x) \ln p_{\boldsymbol{\theta}}(x) - \sum_x p_{\boldsymbol{\theta}}(x) \ln p_{\boldsymbol{\theta}'}(x). \quad (1)$$

注意 KL 散度不对称且不满足三角不等式, 所以严格来说它不是一种“距离”。经典信息几何把分布族  $\{p_{\boldsymbol{\theta}}\}$  看作流形上的点集, 我们仍希望在该流形上定义“距离”。一个自然思路是: 当两分布非常接近时, 用某种距离度量来近似 KL 散度。

考虑参数的微小变动  $\boldsymbol{\theta}' = \boldsymbol{\theta} + \delta\boldsymbol{\theta}$ , 对 KL 散度做泰勒展开 (即展开  $\ln p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}(x)$ ):

$$\begin{aligned}
 D_{\text{KL}}(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}) &= \sum_x p_{\boldsymbol{\theta}}(x) \ln p_{\boldsymbol{\theta}}(x) - \sum_x p_{\boldsymbol{\theta}}(x) \ln p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}(x) \\
 &= \sum_x p_{\boldsymbol{\theta}}(x) \underbrace{[\ln p_{\boldsymbol{\theta}}(x) - \ln p_{\boldsymbol{\theta}}(x)]}_{=0} \quad (\text{零阶项}) \\
 &\quad - \delta\theta^i \underbrace{\sum_x p_{\boldsymbol{\theta}}(x) \partial_i \ln p_{\boldsymbol{\theta}}(x)}_{=0} \quad (\text{一阶项}) \\
 &\quad - \frac{1}{2} \delta\theta^i \delta\theta^j \underbrace{\sum_x p_{\boldsymbol{\theta}}(x) \partial_i \partial_j \ln p_{\boldsymbol{\theta}}(x)}_{:= -I_{ij}(\boldsymbol{\theta})} \quad (\text{二阶项}) \\
 &\quad + O(\|\delta\boldsymbol{\theta}\|^3)
 \end{aligned}$$

其中

- 零阶项为零 (相同分布的 KL 散度为零)
- 一阶项为零: 将概率归一化条件  $\sum_x p_{\boldsymbol{\theta}}(x) = 1$  对  $\theta^j$  求导给出

$$\partial_j \sum_x p_{\boldsymbol{\theta}}(x) \stackrel{\text{求导积分换序}}{=} \sum_x \partial_j p_{\boldsymbol{\theta}}(x) = \sum_x p_{\boldsymbol{\theta}}(x) \partial_j \ln p_{\boldsymbol{\theta}}(x) = 0. \quad (2)$$

即一阶项也为零。这里隐含了一个条件: 对于  $p_{\boldsymbol{\theta}}(x)$ , 求导和积分可以换序, 也即**正则化条件**。

- 二阶项: 对式 (2) 再关于  $\theta^i$  求导,

$$\begin{aligned}
 0 &= \partial_i \sum_x p_{\boldsymbol{\theta}}(x) \partial_j \ln p_{\boldsymbol{\theta}}(x) \\
 &= \sum_x \partial_i [p_{\boldsymbol{\theta}}(x) \partial_j \ln p_{\boldsymbol{\theta}}(x)] \\
 &= \sum_x [p_{\boldsymbol{\theta}}(x) \partial_i \ln p_{\boldsymbol{\theta}}(x) \cdot \partial_j \ln p_{\boldsymbol{\theta}}(x) + p_{\boldsymbol{\theta}}(x) \partial_i \partial_j \ln p_{\boldsymbol{\theta}}(x)].
 \end{aligned}$$

这里隐含的条件是对于  $\partial_j \ln p_{\boldsymbol{\theta}}(x)$ , 求导和积分可以换序。移项得

$$\sum_x p_{\boldsymbol{\theta}}(x) \partial_i \ln p_{\boldsymbol{\theta}}(x) \cdot \partial_j \ln p_{\boldsymbol{\theta}}(x) = - \sum_x p_{\boldsymbol{\theta}}(x) \partial_i \partial_j \ln p_{\boldsymbol{\theta}}(x). \quad (3)$$

将式子左边定义为 **Fisher 信息矩阵**

$$I_{ij}(\boldsymbol{\theta}) := \sum_x p_{\boldsymbol{\theta}}(x) \partial_i \ln p_{\boldsymbol{\theta}}(x) \cdot \partial_j \ln p_{\boldsymbol{\theta}}(x). \quad (4)$$

观察到式子右边是  $\ln p_{\boldsymbol{\theta}}(x)$  对应的 Hessian 矩阵的负期望值, 即

$$\boxed{\text{Fisher 信息矩阵} = -\mathbb{E}_{p_{\boldsymbol{\theta}}} [\text{Hessian of } \ln p_{\boldsymbol{\theta}}(x)]} \quad (5)$$

代入前述 KL 散度的泰勒展开, 得到

$$\boxed{D_{\text{KL}}(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}) \approx \frac{1}{2} \delta\theta^i \delta\theta^j I_{ij}(\boldsymbol{\theta}) = \frac{1}{2} \delta\boldsymbol{\theta}^\top I(\boldsymbol{\theta}) \delta\boldsymbol{\theta}.} \quad (6)$$

这表明在局部我们可以用  $I_{ij}$  作为距离来近似 KL 散度。这是 KL 散度诱导的一个黎曼度规, 称为 **Fisher-Rao 度规**, 从而我们可以将分布们看作一个黎曼流形, 并谈论其上的距离。

表 1: KL 散度泰勒展开各阶项

阶数	值	推导	条件
0	0	$D_{\text{KL}}(p\ p) = 0$	—
1	0	$\partial_i(\sum_x p_{\theta}(x) - 1) = 0$	积分、求导可换序
2	$\frac{1}{2}\delta\theta^{\top}I(\theta)\delta\theta$	$\partial_i\partial_j(\sum_x p_{\theta}(x) - 1) = 0$	积分、求导可换序

## 1.2 Score 向量

为了表述方便，我们定义 **score 向量**（对数似然梯度）

$$s_i(x) := \partial_i \ln p_{\theta}(x),$$

并引入期望记号  $\mathbb{E}_{p_{\theta}}[f(x)] := \sum_x p_{\theta}(x) f(x)$ 。则前述各式可简写为：

- 式 (2):  $\mathbb{E}[s_i] = 0$  (score 均值为零)；
- 式 (4):  $I_{ij} = \mathbb{E}[s_i s_j]$  (Fisher 信息 = score 外积的期望)。

由于  $\mathbb{E}[s_i] = 0$ ,  $I_{ij}$  同时也是 score 的协方差矩阵:  $I_{ij} = \text{Cov}(s_i, s_j)$ 。

**注:** 上述期望取自模型分布  $p_{\theta}$ 。在机器学习中，损失函数的 Hessian  $H_{ij} = \partial_i \partial_j \mathcal{L}$  一般不等于 Fisher 信息矩阵。二者相等当且仅当：(1) 损失为负对数似然  $\mathcal{L} = -\ln p_{\theta}(x)$ ; (2) 期望取自模型分布而非数据分布。

## 1.3 流形上的最速下降：自然梯度

考虑优化损失函数  $\mathcal{L}(\theta)$ , 其梯度为  $g = \nabla_{\theta} \mathcal{L}$ 。普通梯度下降以欧氏范数  $\|\delta\theta\|_2$  约束步长，但这不具备重参数化不变性。若改用 Fisher-Rao 度规  $\|\delta\theta\|_I^2 = \delta\theta^{\top} I(\theta) \delta\theta$  约束步长，则最速下降问题为

$$\max_{\delta\theta} -g^{\top} \delta\theta \quad \text{s.t.} \quad \delta\theta^{\top} I(\theta) \delta\theta = \epsilon^2.$$

用 Lagrange 乘子法，得

$$I(\theta) \delta\theta = -\eta g, \quad \Rightarrow \quad \delta\theta = -\eta I(\theta)^{-1} g. \quad (7)$$

这就是 Amari 提出的**自然梯度** (Natural Gradient Descent, NGD) 更新公式。

# 2 从经典概率分布到量子态：量子 Fisher 信息矩阵

## 2.1 从概率模型类比到量子态

经典概率模型的输出是实非负、归一化的概率分布  $p_{\theta}(x)$ 。量子模型（如参数化变分 Ansatz）的输出则是**复振幅**  $\Psi_{\theta}(x) := \langle x | \Psi(\theta) \rangle$ ，概率通过 Born 规则给出：

$$P(x) = \frac{|\Psi_{\theta}(x)|^2}{\sum_{x'} |\Psi_{\theta}(x')|^2} = \frac{|\Psi_{\theta}(x)|^2}{\langle \Psi | \Psi \rangle}.$$

类比经典的 score 向量  $s_i(x) = \partial_i \ln p_{\theta}(x)$ ，我们定义**对数振幅导数**（复值 score）：

$$O_i(x) := \partial_i \ln \Psi_{\theta}(x) = \frac{\partial_i \Psi_{\theta}(x)}{\Psi_{\theta}(x)}. \quad (8)$$

自然地，尝试直接类比 Fisher 信息矩阵  $I_{ij} = \mathbb{E}_p[s_i s_j]$ ，定义“量子版本”：

$$\tilde{S}_{ij} := \mathbb{E}_P[O_i^* O_j] = \sum_x P(x) O_i^*(x) O_j(x). \quad (9)$$

然而，上述定义  $\tilde{S}_{ij}$  存在根本问题。量子态  $|\Psi\rangle$  乘以任意非零复数  $c$  不改变概率分布： $|c\Psi\rangle \sim |\Psi\rangle$ 。这种规范冗余（全局相位 + 整体幅值）意味着：

1. 若参数变化  $\delta\theta$  只引起  $|\Psi\rangle \rightarrow e^{i\phi}|\Psi\rangle$  这样的全局相位变化，概率分布完全不变，但  $\tilde{S}_{ij}$  却会给出非零值；
2. 类似地，整体归一化变化  $|\Psi\rangle \rightarrow \lambda|\Psi\rangle$  也不影响概率，却被  $\tilde{S}_{ij}$  误计入。

问题的根源在于： $O_i$  中包含了沿  $|\Psi\rangle$  方向的分量，而这一分量对应的正是规范自由度。

解决方案是将  $O_i$  中沿  $|\Psi\rangle$  方向的分量扣除。注意到

$$\mathbb{E}_P[O_i] = \sum_x P(x) \frac{\partial_i \Psi(x)}{\Psi(x)} = \frac{\langle \Psi | \partial_i \Psi \rangle}{\langle \Psi | \Psi \rangle}$$

正是  $O_i$  在  $|\Psi\rangle$  方向的“均值”。定义中心化的对数导数：

$$\bar{O}_i(x) := O_i(x) - \mathbb{E}_P[O_i]. \quad (10)$$

**量子 Fisher 信息矩阵**（又称量子几何张量）定义为其协方差：

$$S_{ij} := \mathbb{E}_P[\bar{O}_i^* \bar{O}_j] = \mathbb{E}_P[O_i^* O_j] - \mathbb{E}_P[O_i^*] \mathbb{E}_P[O_j]. \quad (11)$$

这一定义自动扣除了规范自由度：任何只引起  $|\Psi\rangle \rightarrow c|\Psi\rangle$  的参数变化，其对应的  $\bar{O}_i = 0$ ，从而不贡献  $S_{ij}$ 。

在 braket 符号中， $S_{ij}$  可写为

$$S_{ij} = \frac{\langle \partial_i \Psi | (\mathbb{I} - \Pi_\Psi) | \partial_j \Psi \rangle}{\langle \Psi | \Psi \rangle}, \quad \Pi_\Psi = \frac{|\Psi\rangle \langle \Psi|}{\langle \Psi | \Psi \rangle}, \quad (12)$$

其中投影算符  $\mathbb{I} - \Pi_\Psi$  正是将  $|\partial_i \Psi\rangle$  投影到  $|\Psi\rangle$  的正交补空间。

**注：**在经典概率部分，Fisher 信息矩阵等于负对数似然 Hessian 的期望（式 (4))。但在量子态优化中，损失函数通常是能量期望值  $E(\theta) = \langle \Psi | \hat{H} | \Psi \rangle / \langle \Psi | \Psi \rangle$ ，而非对数似然。因此能量的 Hessian  $\partial_i \partial_j E$  与量子 Fisher 信息矩阵  $S_{ij}$  没有简单的等式关系。

## 2.2 从 Fubini-Study 距离到量子 Fisher 信息矩阵

类比经典的“KL 散度  $\Rightarrow$  Fisher 信息”，量子 Fisher 信息矩阵可从 Fubini-Study (FS) 距离通过二阶展开得到。

对于两个态  $|\Psi\rangle$  与  $|\Phi\rangle$ ，FS 距离定义为

$$d_{\text{FS}}(|\Psi\rangle, |\Phi\rangle) = \arccos \frac{|\langle \Psi | \Phi \rangle|}{\sqrt{\langle \Psi | \Psi \rangle \langle \Phi | \Phi \rangle}}. \quad (13)$$

显然  $d_{\text{FS}}$  对  $|\Psi\rangle \rightarrow c|\Psi\rangle$  不变，因此是定义在射影空间  $\mathbb{CP}^{n-1}$  上的距离。

考虑参数化态  $|\Psi(\theta)\rangle$  与  $|\Psi(\theta + \delta\theta)\rangle$  的 FS 距离。记保真度

$$\mathcal{F}(\delta\theta) = \frac{|\langle \Psi(\theta) | \Psi(\theta + \delta\theta) \rangle|^2}{\langle \Psi(\theta) | \Psi(\theta) \rangle \langle \Psi(\theta + \delta\theta) | \Psi(\theta + \delta\theta) \rangle}.$$

在  $\delta\theta = 0$  处  $\mathcal{F}(0) = 1$ 。将  $\mathcal{F}$  展开到二阶：

$$\mathcal{F}(\delta\theta) \approx 1 - \delta\theta^i \delta\theta^j g_{ij} + O(\|\delta\theta\|^3),$$

其中  $g_{ij} = \text{Re}(S_{ij})$ 。由  $d_{\text{FS}}^2 \approx 2(1 - \sqrt{\mathcal{F}}) \approx g_{ij} \delta\theta^i \delta\theta^j$ ，得到

$$ds_{\text{FS}}^2 = g_{ij} d\theta^i d\theta^j. \quad (14)$$

这正是量子 Fisher 信息矩阵实部  $g_{ij}$  作为 FS 度量的几何意义。

$S_{ij}$  是个厄米的复值张量，实部对称、虚部反对称。将其分解为

$$\boxed{S_{ij} = g_{ij} + \frac{i}{2}F_{ij}, \quad g_{ij} = \text{Re}(S_{ij}), \quad F_{ij} = 2\text{Im}(S_{ij})}. \quad (15)$$

- **实部**  $g_{ij} = g_{ji}$ : 对称半正定，是参数空间上的 **Fubini-Study 度规**，度量两个态的距离；
- **虚部**  $F_{ij} = -F_{ji}$ : 反对称，是 **Berry 曲率**。

### 2.3 流形上的最速下降：随机重构 / 量子自然梯度

类比经典情形，以 FS 度规  $\|\delta\theta\|_g^2 = \delta\theta^\top S(\theta) \delta\theta$  (取  $S$  的实部或对称化版本) 约束步长，最速下降问题为

$$\max_{\delta\theta} -g^\top \delta\theta \quad \text{s.t.} \quad \delta\theta^\top S(\theta) \delta\theta = \epsilon^2.$$

用 Lagrange 乘子法，得

$$\boxed{S(\theta) \delta\theta = -\eta g, \quad \Rightarrow \quad \delta\theta = -\eta S(\theta)^{-1} g}. \quad (16)$$

这就是变分蒙特卡洛中的**随机重构** (Stochastic Reconfiguration, SR)，也称**量子自然梯度** (Quantum Natural Gradient, QNG)。与经典自然梯度  $I(\theta) \delta\theta = -\eta g$  形式完全相同，只是度量从 Fisher 信息矩阵换成了量子 Fisher 信息矩阵。

表 2: 经典概率分布与量子态的对应关系

概念	经典概率模型	量子态模型
参数化对象	概率分布 $p_\theta(x)$	态向量 $ \Psi(\theta)\rangle$ (等价类 $\sim c \Psi\rangle$ )
模型输出	概率 $p(x) \geq 0, \sum_x p = 1$	复振幅 $\Psi(x) \in \mathbb{C}$
规范冗余	仅重参数化	重参数化 + 全局相位 + 整体尺度
对数导数	score $s_i = \partial_i \ln p$	复 score $O_i = \partial_i \ln \Psi$
度量张量	Fisher 信息矩阵	量子 Fisher 信息矩阵
	$I_{ij} = \text{Cov}_p(s_i, s_j)$	$S_{ij} = \text{Cov}_P(O_i^*, O_j)$
度量来源	KL 散度的二阶展开:	FS 距离的二阶展开:
	$D_{\text{KL}} \approx \frac{1}{2}\delta\theta^\top I \delta\theta$	$d_{\text{FS}}^2 \approx \delta\theta^\top \text{Re}S \delta\theta$
额外结构	无 ( $I_{ij}$ 实对称)	Berry 曲率 $F_{ij} = 2\text{Im} S_{ij}$ (反对称)
最速下降	自然梯度 $I \delta\theta = -\eta \nabla \mathcal{L}$	SR/QNG $S \delta\theta = -\eta \nabla E$

## A 优化算法的度量视角

许多优化算法都可以写成“预条件梯度”(preconditioned gradient) 的统一形式：

$$\boxed{M_t \delta\theta_t = -\eta_t \tilde{g}_t, \quad \theta_{t+1} \leftarrow \theta_t + \delta\theta_t}, \quad (17)$$

其中  $t$  是指第  $t$  步， $\tilde{g}_t$  是用于更新的梯度信号 (可以是原始梯度  $\tilde{g}_t = g_t$ ，也可以是动量/偏差校正后的版本)， $M_t \succeq 0$  是预条件矩阵，可视为在参数空间上选取的局部度量：

$$\|\delta\theta\|_{M_t}^2 := \delta\theta^\top M_t \delta\theta.$$

当  $\tilde{g}_t = g_t$  且  $M_t$  可逆时,  $\delta\boldsymbol{\theta}_t = -\eta_t M_t^{-1} g_t$  正对应于在约束  $\|\delta\boldsymbol{\theta}\|_{M_t} \leq \epsilon$  下的最速下降方向; 正文中的自然梯度取  $M_t = I(\boldsymbol{\theta}_t)$ , SR/QNG 取  $M_t = \text{Re } S(\boldsymbol{\theta}_t)$  (或其对称化版本)。若  $M_t$  病态/半正定, 实践中常用阻尼  $M_t \leftarrow M_t + \lambda \mathbb{I}$  来稳定求解。

为避免记号歧义, 以下约定:

- $u \odot v$  表示逐元素乘法,  $u^{\odot 2} := u \odot u$ ;
- 对矩阵  $A$ ,  $\text{diag}(A)$  表示其对角向量,  $\text{Diag}(A)$  表示保留其对角线的对角矩阵;
- 对向量  $v$ ,  $\text{Diag}(v)$  表示以  $v$  为对角元的对角矩阵。

表 3: 优化算法与预条件矩阵  $M_t$  (度量/曲率视角)

方法	预条件/度量 $M_t$	要点 (与几何/曲率的关系)
梯度下降 (SGD)	$M_t = \mathbb{I}$ , $\tilde{g}_t = g_t$	欧氏度量: 各方向同等尺度。
动量法 (Momentum)	$M_t = \mathbb{I}$ , $\tilde{g}_t = m_t$	欧氏度量 + 动量平滑 (对梯度做一阶矩指数平均)。
牛顿法 (Newton)	$M_t = \nabla_{\boldsymbol{\theta}}^2 \mathcal{J}(\boldsymbol{\theta}_t)$ , $\tilde{g}_t = g_t$	用 Hessian 表示局部二阶曲率; 代价高且 $M_t$ 可能不定, 需正则化/截断。
自然梯度 (NG)	$M_t = I(\boldsymbol{\theta}_t)$ , $\tilde{g}_t = g_t$ (Fisher 信息矩阵)	用 KL 散度二阶展开诱导的 Fisher–Rao 度量, 实现重参数化不变的最速下降。
对角自然梯度 (Diag-NG)	$M_t = \text{Diag}(I(\boldsymbol{\theta}_t))$ , $\tilde{g}_t = g_t$	保留 Fisher 的对角线, 忽略参数间相关性; 从 $O(d^2)$ 降到 $O(d)$ 。
AdaGrad	$M_t \approx \text{Diag}(G_t)^{1/2}$ , $G_t = \sum_{\tau \leq t} g_{\tau}^{\odot 2}$	对角自适应缩放: 用历史梯度平方累积来调各维学习率。
RMSProp	$M_t \approx \text{Diag}(v_t)^{1/2}$ , $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^{\odot 2}$	AdaGrad 的指数滑动平均版本 (更适合非平稳训练)。
Adam	$M_t \approx \text{Diag}(\hat{v}_t)^{1/2}$ , $\tilde{g}_t = \hat{m}_t$	对角自适应 + 动量 + 偏差校正。核心: $\hat{v}_t$ 可视为 Fisher 对角线的在线估计, 从而 Adam 等价于“对角 Fisher 预条件”的自然梯度近似。
随机重构 (SR/QNG)	$M_t = \text{Re } S(\boldsymbol{\theta}_t)$ , $\tilde{g}_t = g_t$ (取 $\mathcal{J} = E$ )	以 Fubini–Study 度量为约束得到的最速下降: $M_t \delta\boldsymbol{\theta} = -\eta \nabla_{\boldsymbol{\theta}} E$ 。

**动量 (作为  $\tilde{g}_t$ )** 动量法与 Adam 都引入梯度的一阶矩指数移动平均:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad \hat{m}_t = \frac{m_t}{1 - \beta_1^t}.$$

其中  $\hat{m}_t$  是偏差校正后的版本 (Adam 使用)。在统一形式 (17) 中, 动量对应于保持欧氏度量  $M_t = \mathbb{I}$ , 但把更新方向从  $g_t$  替换为平滑后的  $\tilde{g}_t$ 。

**Fisher 信息矩阵的对角近似 (Adam  $\leftrightarrow$  Diag-NG)** 下面给出把 Adam 与自然梯度联系起来所需的一个最小推导 (与正文 Fisher 定义一致)。

设模型给出概率分布  $p_{\boldsymbol{\theta}}(x)$ , 并采用单样本的负对数似然损失

$$\mathcal{L}(\boldsymbol{\theta}; x) = -\ln p_{\boldsymbol{\theta}}(x), \quad g(\boldsymbol{\theta}; x) = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; x).$$

Fisher 信息矩阵（以模型分布取期望的“true Fisher”）定义为

$$I(\boldsymbol{\theta}) := \mathbb{E}_{x \sim p_{\boldsymbol{\theta}}} [\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(x) \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(x)^{\top}]. \quad (18)$$

由于  $g(\boldsymbol{\theta}; x) = -\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(x)$ , 因此有

$$I(\boldsymbol{\theta}) = \mathbb{E}_{x \sim p_{\boldsymbol{\theta}}} [g(\boldsymbol{\theta}; x) g(\boldsymbol{\theta}; x)^{\top}], \quad \text{diag}(I(\boldsymbol{\theta})) = \mathbb{E}_{x \sim p_{\boldsymbol{\theta}}} [g(\boldsymbol{\theta}; x)^{\odot 2}]. \quad (19)$$

在随机梯度/小批量场景中, 我们用当前梯度  $g_t$  来近似单样本  $g(\boldsymbol{\theta}; x)$ , 并用指数滑动平均来估计逐元素二阶矩:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) (g_t \odot g_t), \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}.$$

因此, 在“负对数似然 + 以梯度平方估计期望”的近似下,

$$\hat{v}_t \approx \text{diag}(I(\boldsymbol{\theta}_t)). \quad (20)$$

这一步就是关键: Adam 选择了 Fisher 信息矩阵的对角近似 (diagonal approximation of FIM), 并用  $\hat{v}_t$  作为其在线估计。

Adam 的完整更新为

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad \hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon},$$

其中除法与开方均为逐元素运算。把它写回统一形式 (17), 可理解为

$$\delta \boldsymbol{\theta}_t = -\alpha \text{diag}(\sqrt{\hat{v}_t} + \epsilon)^{-1} \hat{m}_t,$$

即用一个对角预条件矩阵对梯度 (动量信号) 进行缩放。结合 (20), 可概括为

$$\delta \boldsymbol{\theta}_t \approx -\alpha \text{diag}(I(\boldsymbol{\theta}_t))^{-1/2} \hat{m}_t, \quad (21)$$

这就是“Adam = 对角 Fisher 预条件 (再叠加动量与偏差校正)”的精确表述。

#### 备注: Fisher、经验 Fisher 与 Hessian 的区别

- 自然梯度严格使用的是 (18) 中以  $p_{\boldsymbol{\theta}}$  取期望的 true Fisher; 实践中常用小批量梯度外积近似 (经验 Fisher / empirical Fisher), 它与 true Fisher 一般不完全相同。
- Hessian  $\nabla^2 \mathcal{L}$  与 Fisher  $I(\boldsymbol{\theta})$  也一般不同; 二者在特定条件下才可能一致或近似 (例如匹配模型/大样本/正则性等假设)。

## 参考文献

- [1] S.-i. Amari, *Natural Gradient Works Efficiently in Learning*, Neural Computation **10**(2), 251–276 (1998).
- [2] S. Sorella, *Green Function Monte Carlo with Stochastic Reconfiguration*, Phys. Rev. Lett. **80**, 4558 (1998).
- [3] S. Sorella, *Wave function optimization in the variational Monte Carlo method*, Phys. Rev. B **71**, 241103(R) (2005).
- [4] J. Provost and G. Vallée, *Riemannian structure on manifolds of quantum states*, Commun. Math. Phys. **76**, 289–301 (1980).
- [5] J. Izaac, C. Wang, and Z. Wang, *Quantum Natural Gradient*, arXiv:1811.08451 (2019).