arXiv:2304.05398v1 [math.ST] 10 Apr 2023

基于 Bures–Wasserstein 空间中 JKO 的前向-后向高斯变分推断

Michael Diao* Krishnakumar Balasubramanian† Sinho Chewi‡ Adil Salim§

2023年4月13日

摘要

变分推断（VI）旨在用一个易于处理的分布族中的元素来近似目标分布 π。统计学与机器学习尤其关注高斯变分推断：它通过在高斯分布空间上最小化相对于 π 的 Kullback–Leibler（KL）散度，来近似 π。本文提出（随机）前向-后向高斯变分推断（FB–GVI）算法，以求解高斯变分推断。我们的方法利用了 KL 散度的复合结构：在赋有 Wasserstein 距离的高斯分布 Bures–Wasserstein（BW）空间上，该散度可写成一个光滑项（势函数）与一个非光滑项（熵）之和。当 π 为对数光滑且对数凹时，我们提出的算法可获得最先进的收敛保证；当 π 仅为对数光滑时，我们还首次为其收敛至一阶驻点解提供保证。

1 引言

变分推断（VI）（Blei 等人，2017；Knoblauch 等人，2022年）已成为计算量大的蒙特卡洛马尔可夫链（MCMC）方法的一种易于处理的替代方案。尤其值得关注的是高斯变分推断问题：我们用以下方程的解来近似给定的分布 $\pi \propto \exp(-V)$，其中 V 是一个光滑函数。

$$\operatorname*{arg\,min}_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \mathsf{KL}(\mu \,\|\, \pi)\,, \tag{1}$$

其中 KL 散度 表示 Kullback–Leibler 散度，且 $\mathsf{BW}(\mathbb{R}^d)$ 表示在 $\mathbb{R}^d$ 上的高斯分布 （形式化定义见第 4.1 节）。事实上，高斯变分推断在实践中已展现出更优的性能，尤其是在存在大型数据集的情况下；例如参见Barber and Bishop(1997)；Seeger(1999)；Honkela andValpola (2004年)；Opper and Archambeau (2009)；Quiroz etal. (2022年)。

在关于高斯变分推断的文献中，问题 (1) 的解已被证明具有良好的统计性质，参见例如 Ch´erief-Abdellatif 等人（2019）；Alquier 和 Ridgway（2020）；Katsevich 和 Rigollet（2023）。例如，Katsevich 和 Rigollet（2023）证明，高斯变分推断在估计 π 的均值时优于拉普拉斯近似。此外，考虑 π 是一个足够正则的贝叶斯模型的后验分布的情形。此时，伯恩斯坦 – 冯 · 米塞斯定理（参见 Van der Vaart（2000，第 10 章））以及近期的非渐近结果（Kasprzak 等人，2022年；Spokoiny，2022年）指出，π 可由高斯分布很好地近似，其均值由真实参数的任一渐近有效估计量给出，协方差矩阵则由逆费 Fisher 信息矩阵给出。这些结果共同充分说明了高效计算目标 π 的最佳高斯近似（问题 (1)）具有重要意义。

因此，我们将重点放在优化高斯变分推断的方面，即，求解问题（1）。已有多种方法被提出，我们将在相关工作（第7节）中进行总结。特别是，Lambert 等人（2022b）最近提出了一种用于高斯变分推断的算法，可以将其视为随机梯度下降（法）的类比

*麻省理工学院电气工程与计算机科学系。 diao@mit.edu
†加利福尼亚大学戴维斯分校统计系。 kbala@ucdavis.edu
‡麻省理工学院数学系。 schewi@mit.edu§微软研究院。 adilsalim@microsoft.com

对于问题（1），在赋予 Wasserstein 距离的空间 BW($\mathbb{R}^d$) 上，该空间称为 Bures–Wasserstein（BW）空间。这一观点源于 Wasserstein 空间上的梯度流理论，即，赋予 Wasserstein 距离的分布空间（Jordan 等人，1998年；Ambrosio等人，2008年）。此外，这一观点对于概率推断中的许多问题都发挥了重要作用（相关工作见第 7 节）。然而，从优化的角度来看，Lambert 等人（2022b）所采用的依赖目标泛函 BW 梯度的方法并不是最自然的。事实上，在 BW 空间上，目标泛函 KL（$\cdot \,\|\, \pi$）是复合的：它可以规范地分解为一个称为“势函数”的“光滑”项和一个称为“熵”的“非光滑”项。

KL 散度的复合性质激发了二十多年来关于 Wasserstein 空间上前向-后向方法的研究（例如，Jordan 等人，1998年；Bernton，2018；Wibisono，2018；Salim 等人，2020年）。遗憾的是，这一研究方向受到了所谓 JKO 算子计算难处理性的阻碍（Jordan 等人，1998年）（即，熵在 Wasserstein 空间上的近端算子的对应物）。

本文提出了一种名为（随机）前向-后向高斯变分推断（FB–GVI）的新算法。与Lambert 等人(2022b)类似，BW空间丰富的微分与几何结构是我们方法的关键所在。不过，（随机）FB–GVI 还融入了复合优化和非光滑优化领域中广受认可的思想。本文的一项关键发现是，熵的 JKO 算子限制在 BW空间上时具有闭式解（Wibisono，2018），因此可得出一种可实现的高斯变分推断随机前向-后向（或近端梯度）算法。进而，在多种标准假设下，该算法为高斯变分推断带来了新的最先进计算保证。

我们在下文总结我们的**贡献**。

- 我们提出一种新的（随机）前向-后向算法（随机）FB–GVI，用于求解问题（1）。该算法依赖于熵在 BW空间上的 JKO 算子的闭式公式。
- 我们通过我们的算法，利用近期在概率测度空间上进行优化的技术，证明了高斯变分推断的最先进收敛速率（Ambrosio等人，2008）。

本文其余部分的组织如下。在第 2 节中，我们明确符号并介绍随机和非光滑复合优化的背景知识。在第 3 节中，我们描述 BW 空间的几何与微分结构，这是在 BW($\mathbb{R}^d$) 上进行优化的关键。随后，在第 4 节中，我们聚焦于问题 (1)，并提出我们的算法：FB–GVI 和随机 FB–GVI。我们在第 5 节研究这些算法的收敛性，并在第 6节给出初步的模拟结果。最后，我们在第 7 节讨论相关工作，并在第 8节总结。包含我们实验代码的 Jupyter notebook 可在

https://github.com/mzydiao/FBGVI/blob/main/FBGVI-Experiments.ipynb.

2 背景

本节将说明本文采用的记号，并介绍（随机）前向-后向算法的相关背景。

2.1 记号

我们用$\mathbf{S}^d$ 表示实对称 $d \times d$ 矩阵空间，用$\mathbf{S}^d_{++}$表示实正定 $d \times d$矩阵空间。此外，我们用 I表示 $d \times d$ 维单位矩阵。全文中，$\mathcal{P}_2(\mathbb{R}^d)$表示定义在 $\mathbb{R}^d$ 上的、具有有限二阶矩 $\int \|x\|^2 \mathrm{d}\mu(x) < \infty$的概率测度 μ 集合。令 $\mu \in \mathcal{P}_2(\mathbb{R}^d)$。空间 $L^2(\mu)$是由满足以下条件的 Borel 函数 $f : \mathbb{R}^d \to \mathbb{R}^d$ 构成的希尔伯特空间：

$$\mathbb{E}_\mu \|f\|^2 = \int \|f(x)\|^2 \, \mathrm{d}\mu(x) < \infty \,,$$

配备内积

$$\langle f, g \rangle_\mu \coloneqq \int \langle f(x), g(x) \rangle \, \mathrm{d}\mu(x)$$

及相应的范数 $\|f\|_\mu = \sqrt{\langle f, f\rangle_\mu}$。特别地，恒等映射 $\mathrm{id}: \mathbb{R}^d \to \mathbb{R}^d$ 属于 $L^2(\mu)$。若 $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ 且 $T \in L^2(\mu)$，则 μ 经 T 映射后的推前测度记为 $T_\#\mu$。对于任意可测函数 $\varphi: \mathbb{R}^d \to \mathbb{R}_+$，该推前测度满足 $\int \varphi \, \mathrm{d} T_\#\mu = \int \varphi(T(x)) \, \mathrm{d}\mu(x)$。$\mathcal{P}_2(\mathbb{R}^d)$ 中所有协方差矩阵为正定的高斯分布构成的子集记为 $\mathrm{BW}(\mathbb{R}^d)$。对于元素 $\mu \in \mathrm{BW}(\mathbb{R}^d)$，其均值记为 m_μ，协方差矩阵记为 Σ_μ。记号 $\mathcal{N}(m, \Sigma)$ 表示均值为 $m \in \mathbb{R}^d$、协方差矩阵为 $\Sigma \in \mathbf{S}^d_{++}$ 的高斯分布。

2.2 在 $\mathbb{R}^d$ 上进行随机且非光滑的凸优化

在介绍概率测度空间上的优化概念之前，我们先回顾一下在 $\mathbb{R}^d$ 上进行随机凸非光滑优化的一些细节。首先，如果 $V: \mathbb{R}^d \to \mathbb{R}$ 是 β-光滑的，则 V 二阶连续可微，并且其海森矩阵 $\nabla^2 V(x)$ 在算子范数下被 β 有界，对每个 $x \in \mathbb{R}^d$ 均成立。特别地，V 可微，且其梯度 ∇V 是 β-利普希茨的。此外，V 满足泰勒不等式

$$|V(x+h) - V(x) - \langle \nabla V(x), h\rangle| \le \frac{\beta}{2}\|h\|^2 . \tag{2}$$

考虑以下优化问题

$$\min_{x \in \mathbb{R}^d} \{V(x) + H(x)\}, \tag{3}$$

其中 $V: \mathbb{R}^d \to \mathbb{R}$ 是 β-光滑的，且 $H: \mathbb{R}^d \to \mathbb{R}$ 是凸的，但可能是非光滑的。为简化表述，我们还假设 V 是凸的。由于 H 的非光滑性，将（次）梯度下降算法应用于问题（3）可能不会收敛到 $V + H$ 的极小化器。然而，在许多我们感兴趣的情形下，用户拥有[1] 的近端算子 H 的闭式表达式，定义为

$$\mathrm{prox}_H(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ H(y) + \frac{1}{2}\|x - y\|^2 \right\}. \tag{4}$$

给定近端算子 H 和梯度 V，前向-后向算法（Bauschke 等人，2011）是求解问题（3）最自然、最高效的方法之一。前向-后向算法写作

$$x_{k+1} = \mathrm{prox}_{\eta H}(x_k - \eta \nabla V(x_k)) . \tag{5}$$

在机器学习应用中，用户通常无法直接获得 $\nabla V(x_k)$，因为计算 V 的梯度代价高昂。用户只能获得对 g_k 的一个成本较低的随机估计量 $\hat{} \nabla V(x_k)$。在这种情况下，随机前向-后向算法已被证明是前向-后向算法的一种高效替代方案（Atchad´等人，2017；Bianchi 等人，2019；Gorbunov 等人，2020年）。在随机前向-后向算法中，$\nabla V(x_k)$ 由 $\hat{} g_k$ 替换，如下所示：

$$x_{k+1} = \mathrm{prox}_{\eta H}(x_k - \eta \hat{g}_k) . \tag{6}$$

现在，我们将（随机）前向-后向算法（6）推广到 BW空间。具体而言，第 k 步的迭代量将不再是随机变量 x_k，而是随机高斯分布 p_k。尽管 BW空间并非欧几里得空间，但其固有结构仍使我们能够进行优化，接下来将对此作出说明。

3 Bures–Wasserstein 空间

关于 Wasserstein 空间及其几何学的详细介绍，以及如何据此在该空间中进行优化，可参见 Ambrosio等人 (2008)。本节将简要回顾 BW 空间及其几何学，从而提供在 BW 空间中进行优化并求解问题 (1) 所需的工具。我们首先给出 Wasserstein 空间和 BW 空间的形式化定义。

[1]参见 proximity-operator.net。

3.1 BW空间的几何学

称度量空间 $\mathcal{P}_2(\mathbb{R}^d)$ 配备 2-Wasserstein 距离 W_2 (我们简称为 Wasserstein 距离)后为*Wasserstein* 空间 。回顾一下，对于每个$\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$，Wasserstein 距离定义为

$$W_2^2(\mu,\nu)=\inf_{\gamma\in\mathcal{C}(\mu,\nu)}\int\|x-y\|^2\,\mathrm{d}\gamma(x,y)\,, \tag{7}$$

其中， $\mathcal{C}(\mu,\nu)$是 μ 与 ν之间的耦合集合。BW空间是赋有 Wasserstein 距离 W_2的度量空间 $\mathrm{BW}(\mathbb{R}^d)$。换言之，BW空间是 Wasserstein 空间的一个子集，由所有协方差矩阵为正定矩阵的高斯分布组成。

给定 $\mu,\nu\in\mathrm{BW}(\mathbb{R}^d)$，存在一个从 μ 到 ν的唯一最优传输映射，即，映射 $T\colon\mathbb{R}^d\to\mathbb{R}^d$，使得 $T_{\#}\mu=\nu$ 且

$$W_2^2(\mu,\nu)=\int\|x-T(x)\|^2\,\mathrm{d}\mu(x)\,. \tag{8}$$

换言之，耦合 $(\mathrm{id}\ T)_{\#}\mu$ 属于 $\mathcal{C}(\mu,\nu)$，并在式 (7) 中达到下确界。此外，由于 μ和 ν均为高斯分布， T是一个线性部分为对称矩阵的仿射映射，即可写为 $T(x)=Sx+b$，其中 $S\in\mathbb{S}^d$ 且 $b\in\mathbb{R}^d$ （Olkin 和 Pukelsheim，1982）。特别地，BW空间是一个黎曼流形，在每个 $\mu\in\mathrm{BW}(\mathbb{R}^d)$处，切空间 $\mathfrak{T}_\mu\mathrm{BW}(\mathbb{R}^d)$对应于线性部分为对称矩阵的 d维仿射映射空间。利用 $\mu\in\mathcal{P}_2(\mathbb{R}^d)$， $T_{\#}\mu\in\mathcal{P}_2(\mathbb{R}^d)$ 可推出 $T\in L^2(\mu)$。因此，$\mathfrak{T}_\mu\mathrm{BW}(\mathbb{R}^d)$自然赋有 $L^2(\mu)$内积，从而使 $\mathfrak{T}_\mu\mathrm{BW}(\mathbb{R}^d)$成为 $L^2(\mu)$的有限维子空间。

3.2 BW空间上的优化

本节回顾 BW空间的微分结构。关于 BW空间上的微分演算的更多背景知识见附录A。

3.2.1 一个“光滑”泛函

考虑一个泛函 $\mathcal{F}\colon\mathrm{BW}(\mathbb{R}^d)\to\mathbb{R}$。我们称 $\mathcal{F}$在 μ 处可微，如果存在 $g_\mu\in\mathfrak{T}_\mu\mathrm{BW}(\mathbb{R}^d)$，使得对每个仿射映射 h，

$$\mathcal{F}((\mathrm{id}+th)_{\#}\mu)=\mathcal{F}(\mu)+t\,\langle g_\mu,h\rangle_\mu+o(t)\,. \tag{9}$$

在这种情况下，g_μ 是唯一的，称为 $\mathcal{F}$ 在 μ处的 Bures–Wasserstein 梯度，记作 $\nabla_{\mathrm{BW}}\mathcal{F}(\mu)=g_\mu$。给定一个 β-光滑函数 $V\colon\mathbb{R}^d\to\mathbb{R}$，势能泛函 $\mathcal{V}\colon\mathrm{BW}(\mathbb{R}^d)\to\mathbb{R}$定义为

$$\mathcal{V}(\mu)\coloneqq\int V\,\mathrm{d}\mu \tag{10}$$

for 每个 $\mu\in\mathrm{BW}(\mathbb{R}^d)$ 都是 BW空间上可微泛函的一个有用示例 e.

下一个结果说明势函数是可微的，并给出了其 BW 梯度的公式。BW 梯度的公式可通过对 Lambert 等人(2022b，C.1 节）进行直接改编得到（另见附录A.3）。此外， $\mathcal{V}$ 的可微性在 Wasserstein 空间的相关文献中已有充分论述（Ambrosio等人，2008，定理 10.4.13）；我们将其改编至 BW空间。

引理 3.1 （势函数的 BW 梯度）。考虑式 $\mathcal{V}$ （10）中的势函数泛函，其中 V是 β-光滑的。则， $\mathcal{V}$ 在 μ 处可微，并且以下泰勒不等式成立：对于 h 仿射函数，

$$|\mathcal{V}((\mathrm{id}+h)_{\#}\mu)-\mathcal{V}(\mu)-\langle\nabla_{\mathrm{BW}}\mathcal{V}(\mu),h\rangle_\mu|\le\frac{\beta}{2}\|h\|_\mu^2\,. \tag{11}$$

此外， $\mathcal{V}$ 的 *BW* 梯度有如下闭式表达：

$$\nabla_{\mathrm{BW}}\mathcal{V}(\mu)\colon x\mapsto\mathbb{E}_\mu\nabla V+(\mathbb{E}_\mu\nabla^2V)(x-m_\mu)\,, \tag{12}$$

其中 $m_\mu=\int x\mathrm{d}\mu(x)$ 是 μ的均值。

证明。 方程（12）的证明见Lambert 等人(2022b，C.1 节）（另见附录 A.3），不等式（11）的证明见附录B.1。 □

不等式（11）比可微性更强，可以解释为势函数 $\mathcal{V}$ 在 β-BW空间上是光滑的（注意与不等式（2）的类比）。不等式（11）是 V的光滑性的结果。与 $\mathbb{R}^d$上的优化一样，在处理凸但可能不光滑的泛函时，用户可能更愿意通过其近端算子来处理。

3.2.2 一个“凸”泛函

我们称 $\mathcal{F}$ 是测地凸的，如果对于所有 μ_0， $\mu_1 \in \mathsf{BW}(\mathbb{R}^d)$，

$$\mathcal{F}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{F}(\mu_0), T - \mathrm{id} \rangle_{\mu_0} \leq \mathcal{F}(\mu_1)\,, \tag{13}$$

其中 T 是从 μ_0 到 μ_1的最优传输映射。在这种情况下，我们可以在 $\mathcal{F}$ BW空间上引入近端算子的一个类似物，称为 $\mathcal{F}$ （Jordan 等人，1998年）[2]，其定义为（注意其与（4）的类比）

$$\mathrm{JKO}_{\mathcal{F}}(\mu) \coloneqq \operatorname*{arg\,min}_{\nu \in \mathsf{BW}(\mathbb{R}^d)} \left\{ \mathcal{F}(\nu) + \frac{1}{2} W_2^2(\mu, \nu) \right\}. \tag{14}$$

我们希望强调　　请注意，在我们的定义（14）中，BW JKO 算子定义于 BW空间上　　.

熵 $\mathcal{H}$ 是 BW空间上的一个测地凸泛函的典型例子。更准确地说，熵定义为

$$\mathcal{H}(\mu) = \int \log \mu(x)\, \mathrm{d}\mu(x)\,, \tag{15}$$

对于每个 $\mu \in$BW($\mathbb{R}^d$)，我们将 μ 视为相对于勒贝格测度的密度。

下一个引理指出，熵是测地凸的，并给出了其 BW JKO 算子的公式。BW JKO 算子的公式可通过对 Wibisono(2018，例 7) 的方法作直接改编得到。此外， $\mathcal{H}$ 在 Wasserstein 空间的文献中已被确立为测地凸（Ambrosio等人，2008，注 9.3.10）；我们将此结果改编到 BW空间。

引理 3.2 (熵的 BW JKO)。 考虑熵泛函 $\mathcal{H}$ ，其定义见 (15)。那么， $\mathcal{H}$ 是测地凸的，并且满足以下更强的不等式：对所有 ν， μ_0， $\mu_1 \in \mathsf{BW}(\mathbb{R}^d)$，

$$\mathcal{H}(\mu_0) + \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_{\nu} \leq \mathcal{H}(\mu_1)\,, \tag{16}$$

其中， T_0 （分别为 T_1）是从 ν 到 μ_0 （分别到 μ_1）的最优传输映射。

此外， $\mathcal{H}$ 的 *BW JKO* 算子具有已知的闭式表达：$\mathrm{JKO}_{\eta\mathcal{H}}(\mu)$ （其中 $\eta > 0$）是一个高斯分布，其均值与 μ相同，协方差矩阵为 Σ_1，其中

$$\Sigma_1 = \frac{1}{2}\left(\Sigma + 2\eta I + \left[\Sigma\left(\Sigma + 4\eta I\right)\right]^{1/2}\right), \tag{17}$$

其中， Σ 是 μ的协方差矩阵。

证明。 方程（17）的证明见Wibisono (2018，例 7），不等式（16）的证明见附录B.2。 □

不等式（16）比测地凸性更强，并且是熵的广义测地凸性（Ambrosio等人，2008，注 9.3.10）的推论。最后，（17）给出了熵的 *BW JKO* 闭式解，这一点十分突出，也是我们方法的核心。 作为对比，Salim 等人（2020年）提出了一种算法，它依赖于整个 Wasserstein 空间上的熵的 JKO，但后者的 JKO 无法实现。

[2]在 Jordan 等人 (1998年中，作者将 JKO 算子定义为 Wasserstein 空间上近端算子的类比。受其定义启发，我们定义了 BW空间上的 JKO 算子，并称之为 BW JKO 算子。

4 （随机）前向-后向高斯变分推断

4.1 重新审视高斯变分推断

现在，我们以更正式的方式重述问题（1）。我们假设目标分布 π 关于勒贝格测度具有正密度，也记作 π，这是符号上的一种滥用。我们将 π 写成 $\pi \propto \exp(-V)$ 的形式。此外，我们假设函数 $V\colon\ \mathbb{R}^d \to \mathbb{R}$ 是 β-光滑的。回顾一下，对于任意 $\mu \in \mathsf{BW}(\mathbb{R}^d)$，KL 散度定义为

$$\mathsf{KL}(\mu \,\|\, \pi) = \int \log \frac{\mu(x)}{\pi(x)} \,\mathrm{d}\mu(x)\,. \tag{18}$$

回想一下，我们的目标是求解问题（1）。我们将 $\mathcal{F} := \mathcal{V} + \mathcal{H}$ 定义为势函数（与函数 V相关）和熵之和。简单计算可得，$\mathcal{F}(\mu) - \mathcal{F}(\pi) = \mathsf{KL}\text{ 散度}(\mu \,\|\, \pi)$。由于$\mathcal{F}(\pi)$ 是常数（即，不依赖于 μ），问题（1）等价于

$$\min_{\mu \in \mathsf{BW}(\mathbb{R}^d)} \{\mathcal{V}(\mu) + \mathcal{H}(\mu)\}\,. \tag{19}$$

4.2 所提算法

回想一下，势函数 $\mathcal{V}$ 在 BW空间上是“光滑”的，且 $\mathcal{V}$ 的 BW 梯度具有闭式解（引理 3.1）。还要回想一下，熵 $\mathcal{H}$ 在 BW空间上是“凸”的，且 $\mathcal{H}$ 的 BW JKO 具有闭式解（引理 3.2）。为求解等价问题（19），一个自然的想法是将前向-后向算法推广到 BW空间。这便得到以下前向-后向高斯变分推断（FB–GVI）算法（注意它与（5）的相似之处）：

$$p_{k+\frac{1}{2}} = (\mathrm{id} - \eta\, \nabla_{\mathsf{BW}} \mathcal{V}(p_k))_{\#} p_k\,, \tag{20}$$

$$p_{k+1} = \mathrm{JKO}_{\eta\mathcal{H}}(p_{k+\frac{1}{2}})\,. \tag{21}$$

后向步骤（21）可利用（17）高效求解。尽管前向步骤（20）也有闭式解，但前向步骤需要计算 ∇V 和 $\nabla^2 V$ 关于 p_k的积分，见（12）。这些积分可能难以计算。因此，我们提出通过从 p_k中抽取随机样本，构造 g_k 对 $\nabla_{\mathsf{BW}} \mathcal{V}(p_k)$ 的随机估计，即 即随机梯度。由此得到的算法称为随机 FB–GVI，可写为

$$\begin{aligned} p_{k+\frac{1}{2}} &= (\mathrm{id} - \eta\, \hat{g}_k)_{\#} p_k\,, \\ p_{k+1} &= \mathrm{JKO}_{\eta\mathcal{H}}(p_{k+\frac{1}{2}})\,, \end{aligned} \tag{22}$$

其中，$\hat{}g_k$ 是由下式定义的随机仿射函数

$$\hat{g}_k : x \mapsto \nabla V(\hat{X}_k) + \nabla^2 V(\hat{X}_k)\,(x - m_k)\,, \tag{23}$$

其中，$\hat{}\, X_k$ 从 p_k （即，$X_k \sim p_k$）中抽取，$m_k = \int x\, dp_k(x)$ 是 p_k 的均值。

随机 FB–GVI 是 BW空间上的随机前向-后向算法的一个对应形式（注意它与（6）的类比）。特别地，$(p_k)_{k\in\mathbb{N}}$ 由（22）定义，是一列随机高斯分布，即取值于 $\mathsf{BW}(\mathbb{R}^d)$ 中的随机变量。我们用 m_k表示 p_k 的均值（协方差矩阵），分别用 Σ_k表示。FB–GVI 和随机 FB–GVI 都可以通过迭代序列 p_k的均值和协方差矩阵来实现。FB–GVI 和随机 FB–GVI 关于 m_k 和 Σ_k的迭代见算法 1。可利用为计算矩阵平方根而开发的高效算法（例如，Pleiss 等人 (2020年)；Song 等人 (2022年)）来降低每次迭代的计算复杂度。

5 收敛理论

在本节中，我们将利用 FB–GVI 和随机 FB–GVI 的等价形式（20）–（21）和（22），研究它们的收敛性。我们将使用标准的复杂度记号，例如 $\gtrsim$，$\asymp$。我们还记

算法 1 FB–GVI 和随机 FB–GVI

要求： 步长 $\eta > 0$；迭代次数 N；初始分布 $p_0 = \mathcal{N}(m_0, \Sigma_0)$

对 $k = 0$ **至** $N-1$ **执行**

if FB–GVI**then**

$b_k \leftarrow \mathbb{E}_{p_k} \nabla V$, $S_k \leftarrow \mathbb{E}_{p_k} \nabla^2 V$

否则如果 随机 FB–GVI **那么**

draw $\hat{X}_k \sim \mathcal{N}(m_k, \Sigma_k)$

$b_k \leftarrow \nabla V(\hat{X}_k)$, $S_k \leftarrow \nabla^2 V(\hat{X}_k)$

结束 if

$m_{k+1} \leftarrow m_k - \eta\, b_k$

$M_{k+1} \leftarrow I - \eta\, S_k$

$\Sigma_{k+\frac{1}{2}} \leftarrow M_{k+1} \Sigma_k M_{k+1}$

$\Sigma_{k+1} \leftarrow \frac{1}{2}(\Sigma_{k+\frac{1}{2}}^{-1} + 2\eta I + [\Sigma_{k+\frac{1}{2}}(\Sigma_{k+\frac{1}{2}} + 4\eta I)]^{1/2})$

结束 for

输出 $p_N = \mathcal{N}(m_N, \Sigma_N)$

$\hat{}\pi = \mathcal{N}(\hat{}m, \hat{}\Sigma)$ 是问题（1）的解（即，KL 散度目标的极小化器），并令 $\mathscr{F}_k$ 表示截至第 σ次迭代所生成的 $\hat{}\, k$ -代数（但不包括随机 FB–GVI 中的随机样本 $X_k \sim p_k$ ）。

我们考虑关于 V 的若干假设。给定 $\alpha \in \mathbb{R}$， V是 α-凸的，当且仅当 $\alpha I \preceq \nabla^2 V$。如果 $\alpha = 0$， V被称为凸的；如果 $\alpha > 0$， V被称为（α-）强凸的。对于任一算法，将（随机）误差函数（见算法 b_k 和 S_k 中的定义 1）定义为

$$e_k \colon x \mapsto (S_k - \mathbb{E}_{p_k} \nabla^2 V)(x - m_k) + (b_k - \mathbb{E}_{p_k} \nabla V)\,,$$

并记其期望 $L^2(p_k)$ 范数为 $\sigma_k^2 := \mathbb{E}[\|e_k\|_{p_k}^2 \mid \mathscr{F}_k]$。期望是针对 (b_k, S_k) 可能具有的随机性（即针对 $\hat{}\, X_k$ 的随机性）取的。对于随机 FB–GVI， $e_k = \hat{}g_k - \nabla_{\mathsf{BW}} \mathcal{V}(p_k)$，其中 $\hat{}g_k$ 由 (23) 定义。由于 $\mathbb{E}[e_k \mid \mathscr{F}_k] = 0$ （即 BW 随机梯度是无偏的）， σ_k^2 是第 k 次迭代时 BW 随机梯度的条件方差。对于 FB–GVI， $e_k = \nabla_{\mathsf{BW}} \mathcal{V}(p_k) - \nabla_{\mathsf{BW}} \mathcal{V}(p_k) = 0$，因此$\sigma_k = 0$。我们对 FB–GVI 和随机 FB–GVI 的分析依赖于以下针对 (20)–(21) 和 (22) 两者的迭代序列 $(p_k)_{k \in \mathbb{N}}$ 的统一单步不等式。

引理 5.1 (一步不等式)。 假设 V 是 α-凸且 β-光滑。设 $(p_k)_{k \in \mathbb{N}}$ 为 *FB–GVI* (20)–(21) 或随机 *FB–GVI* (22)的迭代。设 $\eta > 0$ 满足

$$\eta \le \begin{cases} \frac{1}{\beta} & \text{if } \sigma_k = 0 \quad \textit{(FB–GVI)}\,, \\ \frac{1}{2\beta} & \textit{else}\,. \end{cases}$$

然后，对于所有 $\nu \in \mathsf{BW}(\mathbb{R}^d)$，

$$\mathbb{E} W_2^2(p_{k+1}, \nu) \le (1 - \alpha\eta)\, \mathbb{E} W_2^2(p_k, \nu) - 2\eta\, \mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\, \mathbb{E}\sigma_k^2\,. \tag{24}$$

证明。证明见附录C。 □

这一步不等式与随机前向-后向算法分析中经典使用的一步不等式相似，只是将 Wasserstein距离替换为欧几里得距离（Gorbunov 等人，2020年）。引理 5.1 的证明大量运用了第 3节所介绍的 BW空间的微分与几何结构。

5.1 FB–GVI 的收敛

在本节中，$(p_k)_{k \in \mathbb{N}}$ 是由 FB–GVI ((20)–(21)) 定义的迭代序列。令 $\sigma_k = 0$ 取 (24) 中的值，可由引理 5.1 得到推论；当 V 为凸函数或强凸函数时。

定理 5.2 (续 凸情形，FB–GVI)。 假设 V 为凸函数且具有 β-光滑性，并且 $0 < \eta \leq \frac{1}{\beta}$。那么 ，

$$\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \frac{W_2^2(p_0, \hat{\pi})}{2N\eta} .$$

特别地，当 $\eta = \frac{1}{\beta}$和 $N \gtrsim \frac{\beta W_2^2(p_0,\hat{\pi})}{\varepsilon^2}$ 时，我们得到了这一保证。

$$\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2 .$$

证明。 证明见附录 E.1。 □

接下来，我们证明（p_k）收敛至问题（1）的一个极小化器。

定理 5.3 (渐近收敛**)**。 假设 V 是凸的且 β-光滑，并且 $0 < \eta < \frac{1}{\beta}$。那么，问题 的极小化器 $\pi_\star$ (1)存在，使得 $W_2(p_k, \pi_\star) \to 0$ 当 $k \to \infty$。

证明。 证明见附录 E.2。 □

注意，定理 5.3 并非由定理 5.2推出。事实上，定理 5.2 仅能推出序列 $(p_k)_{k\geq 0}$ 弱收敛到问题 ***(1)*** 的极小化器集合。存在唯一的极小化器 $\pi_\star$，使得序列 $(p_k)_{k\geq 0}$ 收敛于该极小化器，这可由在 $\eta < \frac{1}{\beta}$ 处应用一步不等式（引理 5.1），以及 Wasserstein 空间拓扑方面的近期结果（Naldi 和 Savar´e，2021）得出。

定理 5.4 （强凸情形，FB–GVI）。 假设 V 是 α-强凸且 β-光滑，并且 $0 < \eta \leq \frac{1}{\beta}$。那么，

$$W_2^2(p_N, \hat{\pi}) \leq \exp(-\alpha N\eta)\, W_2^2(p_0, \hat{\pi}) .$$

特别地，当 $\eta = \frac{1}{\beta}$ 且 $N \gtrsim \frac{\beta}{\alpha}$ 对数 $\frac{W_2(p_0,\hat{\pi})}{\varepsilon}$ 时，我们得到如下保证

$$\alpha\, W_2^2(\mu_N, \hat{\pi}) \leq \varepsilon^2 , \quad \textit{and} \quad \mathcal{F}(p_{2N}) - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2 .$$

证明。 证明见附录 E.3。 □

定理 5.2 指出，对于凸 V，FB–GVI 具有次线性收敛性（就目标函数差距而言）；定理 5.4 指出，对于强凸 V，FB–GVI 具有线性收敛性。其收敛速率与确定性前向-后向算法在 $\mathbb{R}^d$ （Gorbunov 等人，2020年）上的收敛速率同阶。

最后，我们还将结果推广到非凸情形，并得到了驻点保证。

定理 5.5 （非凸情形，FB–GVI）。 假设 V 是 β-光滑的，且 $0 < \eta \leq \frac{1}{\beta}$。令$\Delta := \mathcal{F}(p_0) - \mathcal{F}(\hat{\pi})$。那么，

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|_{p_k}^2 \leq \frac{150\Delta}{\eta N} .$$

特别地，当 $\eta = \frac{1}{\beta}$ 且 $N \gtrsim \frac{\beta\Delta}{\varepsilon^2}$ 时，我们得到如下保证

$$\min_{k\in\{0,\dots,N-1\}} \|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|_{p_k}^2 \leq \varepsilon^2 . \tag{25}$$

证明。 证明见附录E.4。 □

据我们所知，这是高斯变分推断的首个驻点保证。该结果的意义在于，根据Katsevich 和 Rigollet （2023年），高斯变分推断的优良统计性质并非源于对（1）中目标函数的全局最小化，而是源于一阶最优性条件（25）。因此，即使在非对数凹情形下，定理 5.5也可视为关于后验分布近似的算法性结果。

我们还要指出，在定理 5.5 中，尽管我们假设 V 是光滑的，但这并不意味着目标函数 $\mathcal{F}$ 在 Bures–Wasserstein空间上是光滑的，因为其中含有熵项 $\mathcal{H}$。事实上，定理 5.5 的证明需要对 FB–GVI 迭代序列的特征值进行仔细控制。

5.2 随机前向-后向广义变分推断的收敛

在本节中，$(p_k)_{k\in\mathbb{N}}$ 是由 (22) 定义的迭代序列。为了使用引理 5.1，我们先证明 σ_k^2 的一个上界，即 BW 随机梯度的方差。

引理 5.6. 如果 V 是凸且 β-光滑的，那么

$$\sigma_k^2 \le 6\beta d + 12\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(p_k, \hat{\pi})\,.$$

此外，若 V 是 α-强凸的，上述界变为

$$\sigma_k^2 \le 6\beta d + \frac{12\beta^3}{\alpha}\, W_2^2(p_k, \hat{\pi})\,.$$

证明。见附录F.1。 □

关于 σ_k^2 的界令人联想到随机梯度算法文献中对 $\mathbb{R}^d$常见的假设，即随机梯度具有次线性增长（Kushner 和 Yin，2003；Bottou 等人，2018）。我们强调，我们并未假设这种次线性增长。相反，引理 5.6 证明了随机 FB–GVI 中使用的 BW 随机梯度具有次线性增长。接下来，通过利用引理 5.1控制 σ_k^2，我们推导出随机 FB–GVI 的推论。引理 5.6。

定理 **5 7** V β $0 < \eta \le$。(凸情形，随机 FB–GVI)。 $\frac{1}{2\beta}$ $c = 24\beta^3\lambda_{\max}\ \Sigma$
假设是凸的且是-光滑的，并且。定义:(ˆ)。那么，

$$\mathbb{E}\big[\min_{k\in\{1,\dots,N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \le \frac{2W_2^2(p_0,\hat{\pi})}{N\eta} + 2c\eta\, W_2^2(p_0,\hat{\pi}) + 12\beta\eta d\,.$$

特别地，对于足够小的 ε^2/d 值且

$$\eta \asymp \frac{\varepsilon^2}{cW_2^2(p_0,\hat{\pi}) \vee \beta d}\,, \quad \textit{and} \quad N \gtrsim \frac{W_2^2(p_0,\hat{\pi})}{\varepsilon^4}\left(cW_2^2(p_0,\hat{\pi}) \vee \beta d\right),$$

我们获得了这一保证。

$$\mathbb{E}\big[\min_{k\in\{1,\dots,N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \le \varepsilon^2\,.$$

证明。 参见附录F.3。 □

定理5.8 (强凸情形，随机 FB–GVI)。假设 V 是 α-强凸的，且 β-光滑，并且 $\eta \le \frac{\alpha^2}{48\beta^3}$。那么，

$$\mathbb{E}W_2^2(p_N,\hat{\pi}) \le \exp\Big(-\frac{\alpha N\eta}{2}\Big)\, W_2^2(p_0,\hat{\pi}) + \frac{24\beta\eta d}{\alpha}\,.$$

特别地，当 ε^2/d 足够小时，并且

$$\eta \asymp \frac{\varepsilon^2}{\beta d}\,, \quad \textit{and} \quad N \gtrsim \frac{\beta d}{\alpha\varepsilon^2}\log\frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2}\,,$$

我们得到了这些保证。

$$\alpha\,\mathbb{E}W_2^2(p_N,\hat{\pi}) \le \varepsilon^2\,, \quad \textit{and} \quad \mathbb{E}\big[\min_{k\in\{1,\dots,2N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \le \varepsilon^2\,.$$

证明。参见附录F.4。 □

据我们所知，定理 5.7 是首个针对对数光滑的对数凹目标分布，就问题（1）中的目标函数差距给出复杂度结果的研究结果。此外，定理 5.8 改进了 Lambert 等人 （2022b）针对强对数凹目标分布所获得的当前最佳结果。特别地，忽略对数因子后，他们的迭代复杂度（以尺度不变的方式表示时）为 $\widetilde{O}\left(\frac{\beta^2 d}{\alpha^2\varepsilon^2}\right)$，而我们的迭代复杂度为 $\widetilde{O}\left(\frac{\beta d}{\alpha\varepsilon^2}\right)$。注意，条件数 β/α的线性依赖关系对于梯度下降方法而言是意料之中的。我们指出，我们的分析关键性地利用了非光滑熵上的近端算子（BW JKO），从而获得了改进的收敛率。

6 仿真

在本节中，通过基础模拟[3],我们证明了 FBGVI 是可实现、实用且具有竞争力的，其可与Lambert 等人 （2022b）提出的Bures–Wasserstein 梯度下降（BWGD）方法相媲美。我们考虑两个示例：

1. 高斯分布目标。对于第一个实验，我们考虑目标密度为 $\pi(x) \propto \exp(-\frac{1}{2}\langle (x-\mu)\ \Sigma^{-1}(x-\mu)\rangle)$ 的情形，其中 $\mu \sim \mathrm{Unif}([0,1]^{10})$ 且 $\Sigma^{-1} = U\,\mathrm{diag}\begin{bmatrix}10^{-9} & 10^{-8} & \cdots & 1\end{bmatrix}U^{\mathsf{T}}$，$U \in \mathbb{R}^{10\times 10}$ 选为均匀随机正交矩阵。在这种情况下，我们有 $\pi \in \mathrm{BW}(\mathbb{R}^{10})$，因此问题 (1) 的解恰好是 π；此外，我们有 π 是 10^{-9}-强对数凹且 1-对数光滑的。

   我们运行 FB–GVI 和随机 FB–GVI，目标势函数为 $\pi \propto \exp(-V)$，初始化为 $p_0 = \mathcal{N}(0, I_{10})$，其中 I_{10} 是 10×10 单位矩阵。步长 η 不断变化，不同 $\mathrm{KL}(p_k \,\|\, \pi$取值对应的 log η 曲线如图 1所示。

2. **贝叶斯逻辑回归**。 我们考虑以下生成模型：给定参数 $\theta \in \mathbb{R}^d$，我们抽取独立同分布样本 $\{(X_i, Y_i)\}_{i=1}^n \in (\mathbb{R}^d \times \{0,1\})^n$ 其中

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)\,, \quad Y_i \mid X_i \sim \mathrm{Bern}(e^{\langle \theta, X_i\rangle})\,.$$

   给定这些样本 $\{(X_i, Y_i)\}_{i=1}^n$ 以及关于 θ 的均匀（非正规）先验， θ 的后验分布为

$$V(\theta) = \sum_{i=1}^{n} \left[\ln(1 + e^{\langle \theta, X_i\rangle}) - Y_i \langle \theta, X_i\rangle\right].$$

   我们运行随机 FB–GVI，并使用 $\pi \propto \exp(-V)$，以 $p_0 = \mathcal{N}(0, I_d)$ 为初始值，同时改变步长 η。由于在此情形下，我们不知道真实的极小化器 $\hat{}\pi$ ，也不知道 π 的归一化常数，因此无法直接计算 $\mathrm{KL}(p_k \,\|\, \pi)$ 或 $W_2^2(p_k, \hat{}\pi)$。不过，我们仍可以通过从 p_k 抽取样本，经验估计目标函数 $\mathcal{F}(p_k)$ 以及 BW 梯度范数平方 $\mathbb{E}_{p_k}\|\nabla_{\mathrm{BW}}\mathcal{F}(p_k)\|^2$ 。对于每个步长 η，我们在图 2 中绘制了 $\mathcal{F}(p_k)$ 和 $\mathbb{E}_{p_k}\|\nabla_{\mathrm{BW}}\mathcal{F}(p_k)\|^2$ 随迭代次数的经验估计值。

我们的结果见图 1 和图 2。根据图示，我们作出以下观察：

1. （随机）FB–GVI 的表现与 BWGD 一样好，甚至更好。此外，当 η足够小时，FB–GVI 和 BWGD 都能达到比拉普拉斯近似更低的目标值，如图 2 所示。Katsevich 和 Rigollet Katsevichand Rigollet （2023年）也对 BWGD 做出了这一观察。

2.FB–GVI的稳定性可维持到远大于 BWGD的步长，这与欧几里得空间中近端梯度方法相较于梯度下降方法的稳定性优势相呼应，尤其是在步长较大时（参见例如，Toulis 等人（2016年，2021））。

我们的实证结果结合理论保证，为使用 FB–GVI 进行高斯变分推断提供了有力证据。

7 相关工作

现在我们讨论与本文密切相关的研究脉络，并将本文置于更广泛的采样和变分推断文献背景中。

[3]我们实验的代码可在 https://github.com/mzydiao/FBGVI/blob/main/FBGVI-Experiments.ipynb找到。

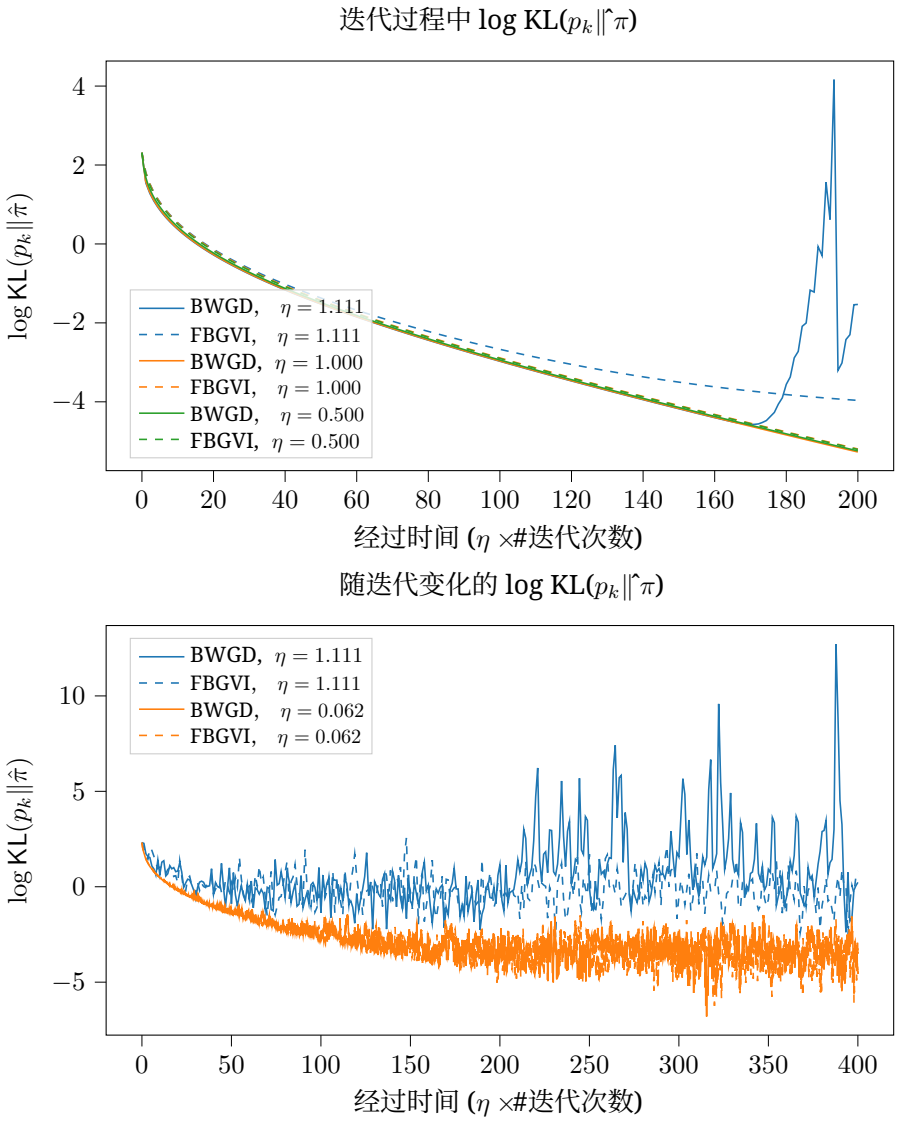

图 1：高斯目标实验：FB–GVI（上）和随机 FB–GVI（下）的结果。

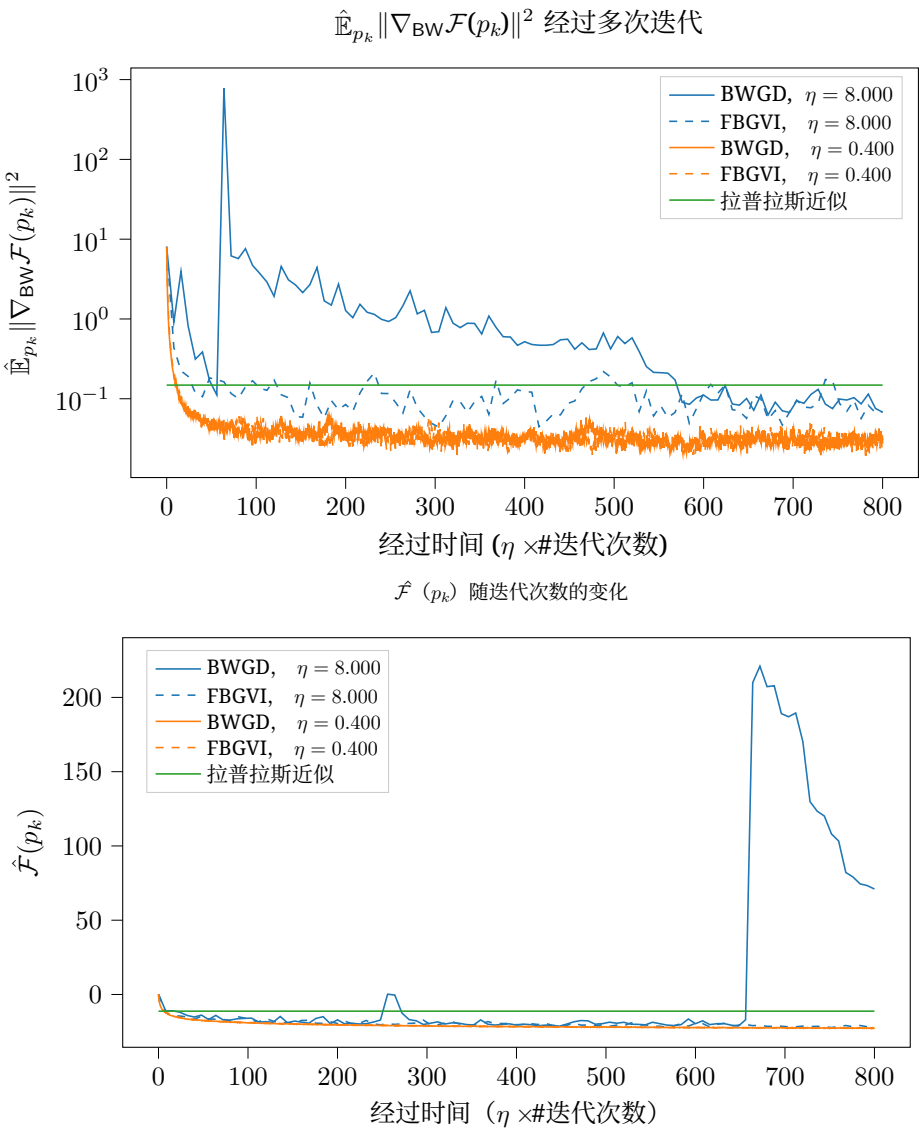

图 2：贝叶斯逻辑回归实验：随机 FB–GVI 的 $\log \hat{\mathbb{E}}_{p_k}\|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|^2$（上）和 $\hat{\mathcal{F}}(p_k)$（下）的图。

高斯变分推断的优化算法。Paisley 等（2012）；Ranganath 等（2014）；Lambert 等（2022a）研究了用于求解高斯变分推断的算法。一般方法是对高斯分布集合进行参数化，并应用欧几里得优化。具体而言，Alquier 和 Ridgway（2020）注意到，当 π 是贝叶斯逻辑回归中的后验分布时，经过某种特定的参数化，问题 (1) 会变为凸问题。在这种情况下，他们还刻画了梯度下降迭代点的统计性质。Challis 和 Barber（2013）；Domke（2020）还给出了其他相应优化问题为凸问题的情形。具体而言，Domke（2020）表明，在采用 Alquier 和 Ridgway（2020）的参数化方式时，模型负对数密度在欧几里得空间中的光滑性／凸性性质，会使其在参数空间中具有相同的光滑性／凸性性质。然而，要获得随机情形下的收敛速率，需要控制随机梯度的方差。这项并非易事的工作并未在 Domke（2020）中完成。在我们的工作中，所需的方差控制由引理 5.6 建立，并且是获得收敛速率的关键一步。

基于自然梯度方法（Zhang 等人，2018；Lin 等人，2019、2020年）和归一化流（Rezende 和 Mohamed，2015；Kingma 等人，2016年；Caterini 等人，2021也已被提出用于变分推断。然而，据我们所知，文献中尚无这类方法的收敛结果。

最后，在相关文献中，与我们工作最接近的是Lambert 等人（2022b），他们同样提出了一种在BW空间上进行优化的算法，称为 Bures–Wasserstein 随机梯度下降（BW–SGD），以求解问题（1）。他们的算法 BW–SGD 依赖于对非光滑熵取梯度；具体而言，他们仅能在 π 为强对数凹时给出一个（次优的）收敛速率。在本工作中，我们不仅改进了强对数凹情形下的收敛速率，还证明了对数凹情形下也具有收敛速率。

在 Wasserstein 空间上最小化 KL 。正如前文所述，我们的方法源于近期文献中的一种观点，即将采样方法视为 Wasserstein 空间上的优化算法。

例如，Langevin蒙特卡洛（LMC）算法（Dalalyan，2017）是一种用于从目标分布 π 中进行采样的马尔可夫链蒙特卡罗算法。Wasserstein 梯度流理论（Ambrosio等人，2008）提供了相应的数学工具，使我们能够将 LMC（及其众多变体）视为 Wasserstein 空间上的优化算法。对于 LMC，要最小化的目标函数是 $\mathrm{KL}(\cdot \,\|\, \pi)$。因此，可以使用优化分析（在 Wasserstein 空间上）来证明 LMC 的收敛界（Wibisono，2018；Durmus 等人，2019；Balasubramanian 等人，2022年；Chen 等人，2022年；Chewi，2023年）。

Stein变分梯度下降（Liuand Wang, 2016年；Liu, 2017）是另一种可被视为用于最小化 KL散度$(\cdot \,\|\, \pi)$的优化算法。SVGD 是一种确定性算法，它驱动一组粒子的经验分布以拟合 π。SVGD的迭代通过反复应用一个精心选择的映射 T 来计算，该映射使得 $T - I$ 属于再生核希尔伯特空间（RKHS）。关于SVGD 的收敛速率，我们所知甚少（Duncan等人，2019；Lu等人，2019；Chewi等人，2020a；Korba 等人，2020；Salim 等人，2022年；He 等人，2022年；Shi and Mackey, 2022）。然而，SVGD 与 BW–GD 之间存在一个有趣的联系（Lambert等人，2022b）：当 SVGD的粒子数量趋于无穷时（“均值场”极限），如果 RKHS是由线性部分对称的仿射函数构成的集合，则 BW–GD 的迭代等价于 SVGD 的迭代。我们还要指出，已有其他用于在 Wasserstein空间上进行基于粒子的优化的启发式算法被提出，例如 Carrillo 等人（2019）；Alvarez-Melis 等人（2022）；Backhoff-Veraguas 等人（2022）；Wang 等人（2022a）；Yao and Yang（2022），但这些算法都没有任何非渐近收敛保证。

与我们的工作密切相关的另一项工作是 Salim 等人 (2020年。与我们的工作类似，他们将目标函数 KL 散度$(\cdot \,\|\, \pi)$ 视为 Wasserstein 空间上的复合泛函。他们提出了一种前向-后向算法，其中包含熵的 JKO 步骤，并具有强收敛性质。然而，他们没有讨论熵的 JKO 步骤如何实现。因此，据我们所知，他们的算法无法实现。相反，我们的算法依赖于 BW 空间上的熵的 JKO 步骤，而该步骤已被证明存在闭式解。

（非光滑）流形优化。 我们的工作在理念上也与近期一些研究相关，这些研究致力于开发高效算法，以求解特定流形上的非光滑优化问题。例如，在确定性设定下，Li 等人 (2021) 分析了次梯度方法，Chen 等人(2020)；Huang 和 Wei (2022) 分析了近端梯度方法，Chen 等人 (2021) 分析了近端点方法，

和Wang 等人(2022b) 分析了近端线性方法。随机版本在Li 等人(2022年)；Wang 等人(2022b) 中进行了研究。我们还可参阅Zhang 等人(2021)；Hu 等人(2022年)；Peng 等人(2022年)；Zhang 和 Davanloo Tajbakhsh(2022年)，了解非光滑流形优化的其他近期进展。上述多项工作考虑了黎曼流形这一一般情形。BW 空间是一个黎曼流形，同时也是 Wasserstein 空间的一个子集；我们的工作正是利用了这一结构来证明收敛结果。

BW空间的几何学已在Modin (2017) 中得到研究；Malag`o 等人 (2018)；Bhatia 等人(2019)，在该空间上进行优化已被证明对各种应用大有裨益，例如参见Chewi 等人 (2020b)；Altschuler 等人 (2021)；Han 等人 (2021)；Lambert 等人 (2022b)；Luo 和 Trillos(2022)；Maunu 等人(2022)。

8 结论

我们提出了一种新颖的优化算法，即用于求解（1）中的高斯变分推断问题的（随机）FB–GVI。我们将该算法视为在Bures–Wasserstein空间上执行优化，这与一系列关于受优化启发的采样和变分推断算法设计与分析的成功研究相呼应。基于这一视角，我们还针对求解（1）给出了新的或当前最优的收敛速率，具体取决于 π所满足的正则性假设。作为近期的直接后续工作，研究所提出的（随机）FW–GVI算法在各种受关注的具体实际问题上的统计性质（一致性、正态近似界、矩估计界和稳健性）颇具意义。

从更广泛的层面来看，我们的工作为以下问题打开了大门：我们能否为一般的 VI开发一个严格的算法框架，即，问题（1）其中 $\mathsf{BW}(\mathbb{R}^d)$ 被不同的或更大的分布集合所取代（例如，高斯分布的混合）？我们相信，本文为实现这一总体目标迈出了切实的一步。

致谢

我们感谢匿名审稿人建议加入定理 5.3 及其证明，并感谢其让我们注意到Domke(2020年) 的工作。KB部分获得了美国国家科学基金会（NSF）DMS-2053918 号资助。SC 获得了 NSF TRIPODS项目（资助编号 DMS-2022448）的支持。

参考文献

K. Ahn和S.Chewi。通过镜像 Langevin 算法进行高效的约束采样。载于 M. Ranzato、A. Beygelzimer、K. Nguyen、P. S. Liang、J. W.Vaughan 和 Y. Dauphin 编，神经信息处理系统进展，第 34 卷，第28405–28418 页。Curran Associates, Inc.，2021。（第23页引用。）P.Alquier 和J.Ridgway。温度化后验分布及其变分近似的集中性。统计学年鉴，48(3):1475–1497，2020年。（第 1 页和第13 页引用。）J. Altschuler、S. Chewi、P.R. Gerber 和 A.Stromme。在Bures–Wasserstein流形上取平均：梯度下降的无维收敛性。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P. Liang 和 J. W. Vaughan 编，神经信息处理系统进展，第 34 卷，第 22132–22145 页。Curran Associates, Inc.，2021。（第 14、23 和 25 页引用。）D. Alvarez-Melis、Y. Schiff和 Y. Mroueh。使用输入凸神经网络优化概率空间上的泛函。机器学习研究汇刊，2022年。（第 13 页引用。）L.Ambrosio、N. Gigli 和G.Savar´e。梯度流：度量空间和概率测度空间中的梯度流。SpringerScience& Business Media，2008。（第 2、3、4、5、13、20 和 23 页引用。）Y. F. Atchad´e、G. Fort 和E. Moulines。关于扰动的近端梯度算法。机器学习研究杂志，18(1):310–342，2017。（第 3 页引用。）

J. Backhoff-Veraguas、J. Fontbona、G. Rios 和 F. Tobar。Wasserstein 空间中的随机梯度下降。arXiv 预印本 arXiv:2201.04232，2022年。（引于第13页。）K. Balasubramanian、S. Chewi、M. A. Erdogdu、A. Salim 和 S. Zhang。迈向非对数凹采样理论：Langevin蒙特卡洛的一阶驻点保证。载于《学习理论会议》，第2896–2923页。PMLR，2022年。（引于第13页。）D. Barber 和 C. Bishop。多层网络的集成学习。神经信息处理系统进展，10，1997年。（引于第1页。）H. H. Bauschke、P. L. Combettes 等。希尔伯特空间中的凸分析与单调算子理论，第408卷。Springer，2011年。（引于第3页。）E. Bernton。Langevin蒙特卡洛与JKO分裂。载于 S. Bubeck、V. Perchet 和 P. Rigollet 编，《第31届学习理论会议论文集》，机器学习研究论文集第75卷，第1777–1798页。PMLR，2018年7月6–9日。（引于第2页。）R. Bhatia、T. Jain 和 Y. Lim。正定矩阵之间的 Bures–Wasserstein 距离。Expo. Math.，37(2):165–191，2019年。（引于第14页。）P. Bianchi、W. Hachem 和 A. Salim。涉及随机极大单调算子的常步长前向-后向算法。J. Convex Anal.，26(2):397–436，2019年。（引于第3页。）D. M. Blei、A. Kucukelbir 和 J. D. McAuliffe。变分推断：统计学家综述。Journal of the American Statistical Association，112(518):859–877，2017年。（引于第1页。）L. Bottou、F. E. Curtis 和 J. Nocedal。大规模机器学习的优化方法。《SIAM评论》，60(2):223–311，2018年。（引于第9页。）H. J. Brascamp 和 E. H. Lieb。关于 Brunn–Minkowski 和 Pr´ekopa–Leindler 定理的扩展，包括对数凹函数的不等式及其在扩散方程中的应用。J. Functional Analysis，22(4):366–389，1976年。（引于第34页。）J. A. Carrillo、K. Craig 和 F. S. Patacchini。一种用于扩散的 blob 方法。《变分法与偏微分方程》，58:1–53，2019年。（引于第13页。）A. Caterini、R. Cornish、D. Sejdinovic 和 A. Doucet。使用连续索引归一化流进行变分推断。载于《人工智能中的不确定性》，第44–53页。PMLR，2021年。（引于第13页。）E. Challis 和 D. Barber。高斯 Kullback–Leibler 近似推断。《机器学习研究杂志》，14(8)，2013年。（引于第13页。）S. Chen、S. Ma、A. Man-Cho So 和 T. Zhang。Stiefel流形上的非光滑优化近端梯度法。《SIAM优化杂志》，30(1):210–239，2020年。（引于第13页。）S. Chen、Z. Deng、S. Ma 和 A. M.-C. So。用于对偶主成分追踪和正交字典学习的流形近端点算法。IEEE Transactions on Signal Processing，69:4759–4773，2021年。（引于第13页。）Y. Chen、S. Chewi、A. Salim 和 A. Wibisono。用于采样的近端算法的改进分析。载于《学习理论会议》，第2984–3014页。PMLR，2022年。（引于第13页。）B.-E. Ch´erief-Abdellatif、P. Alquier 和 M. E. Khan。在线变分推断的一个泛化界。载于《亚洲机器学习会议》，第662–677页。PMLR，2019年。（引于第1页。）S. Chewi。对数凹采样。2023年。见 https://chewisinho.github.io/。（引于第13页。）S. Chewi、T. Le Gouic、C. Lu、T. Maunu 和 P. Rigollet。SVGD 是卡方散度的核化 Wasserstein 梯度流。神经信息处理系统进展，33:2098–2109，2020a。（引于第13页。）

S. Chewi、T. Maunu、P.Rigollet 和 A.J. Stromme。《Bures–Wasserstein重心的梯度下降算法》。载于 J.Abernethy 和S. Agarwal（编），《第三十三届学习理论会议论文集》，《机器学习研究论文集》第125卷，第1276–1304页。PMLR，2020年7月9日至12日。(引文见第14、23和24页。) A. S. Dalalyan。《从光滑对数凹密度进行近似采样的理论保证》。《英国皇家统计学会会刊：B辑（统计方法学）》，79(3):651–676，2017。(引文见第13页。) J. Domke。《黑箱变分推断的可证明光滑性保证》。载于《国际机器学习会议》，第2587–2596页。PMLR，2020年。(引文见第13和14页。) A. Duncan、N. N ̈usken 和 L. Szpruch。《关于Stein变分梯度下降的几何学》。arXiv预印本arXiv:1912.00894，2019。(引文见第13页。) A. Durmus、S. Majewski 和B.Miasojedow。《通过凸优化分析Langevin蒙特卡洛》。《机器学习研究杂志》，20(1):2666–2711，2019。(引文见第13和37页。) E.Gorbunov、F.Hanzely和P.Richt ́arik。《随机梯度下降的统一理论：方差缩减、采样、量化和坐标下降》。载于《国际人工智能与统计会议》，第680–690页。PMLR，2020年。(引文见第3、7和8页。) A.Han、B. Mishra、P. Jawanpuria和J. Gao。《具有Bures–Wasserstein几何的正定矩阵上的黎曼优化》。载于 A. Beygelzimer、Y. Dauphin、P. Liang 和 J.W.Vaughan（编），《神经信息处理系统进展》，2021年。(引文见第14页。) Y. He、K. Balasubramanian、B. K.Sriperumbudur和 J. Lu。《正则化Stein变分梯度流》。arXiv预印本arXiv:2211.07861，2022年。(引文见第13页。)A.Honkela 和H. Valpola。《非线性模型的无监督变分贝叶斯学习》。《神经信息处理系统进展》，17，2004年。(引文见第1页。) X. Hu、N. Xiao、X. Liu和 K.-C.Toh。《Stiefel流形上非光滑优化的约束消解方法》。arXiv预印本arXiv:2205.10500，2022年。(引文见第14页。) W. Huang和 K. Wei。《黎曼近端梯度法》。《数学规划》，194(1-2): 371–413，2022年。(引文见第13页。) R. Jordan、D. Kinderlehrer和 F. Otto。《Fokker–Planck方程的变分形式》。《SIAM数学分析杂志》，29(1):1–17，1998年。(引文见第2和5页。) M. J. Kasprzak、R.Giordano 和T.Broderick。《你的后验分布高斯近似有多好？针对多种有用散度的有限样本可计算误差界》。arXiv预印本arXiv:2209.14992，2022年。(引文见第1页。)A.Katsevich 和 P.Rigollet。《关于高斯变分推断的近似精度》。arXiv预印本arXiv:2301.02168，2023年。(引文见第1、8和10页。) D. P. Kingma、T.Salimans、R. Jozefowicz、X. Chen、I.Sutskever和 M. Welling。《利用逆自回归流改进变分推断》。《神经信息处理系统进展》，29，2016年。(引文见第13页。) J. Knoblauch、J.Jewson和 T.Damoulas。《以优化为中心看待贝叶斯定理：回顾并推广变分推断》。《机器学习研究杂志》，23(132):1–109，2022年。(引文见第1页。) A. Korba、A. Salim、M. Arbel、G.Luise和 A. Gretton。《Stein变分梯度下降的非渐近分析》。《神经信息处理系统进展》，33:4672–4682，2020年。(引文见第13页。)

H. Kushner 和 G. Yin。随机逼近与递归算法。载于《随机建模与应用概率》，第 35 卷。Springer-Verlag NY，2003。（引用见第 9 页。）M. Lambert、S. Bonnabel 和 F. Bach。递归变分高斯近似（R-VGA）。《统计与计算》，32(1):10，2022a。（引用见第 13 页。）M. Lambert、S. Chewi、F. Bach、S. Bonnabel 和 P. Rigollet。通过 Wasserstein 梯度流进行变分推断。载于 A. H. Oh、A. Agarwal、D. Belgrave 和 K. Cho 编，《神经信息处理系统进展》，2022b。（引用见第 1、2、4、5、9、10、13、14、20、21 和 22 页。）J. Li、K. Balasubramanian 和 S. Ma。随机零阶黎曼导数估计与优化。《运筹学数学》，2022。（引用见第 14 页。）X. Li、S. Chen、Z. Deng、Q. Qu、Z. Zhu 和 A. Man-Cho So。利用黎曼次梯度型方法在 Stiefel 流形上进行弱凸优化。《SIAM优化杂志》，31(3):1605–1634，2021。（引用见第 13 页。）W. Lin、M. E. Khan 和 M. Schmidt。使用指数族近似混合的快速简易自然梯度变分推断。载于《国际机器学习会议》，第 3992–4002 页。PMLR，2019。（引用见第 13 页。）W. Lin、M. Schmidt 和 M. E. Khan。在贝叶斯学习规则中处理正定约束。载于《国际机器学习会议》，第 6116–6126 页。PMLR，2020。（引用见第 13 页。）Q. Liu。作为梯度流的 Stein 变分梯度下降。《神经信息处理系统进展》，30，2017。（引用见第 13 页。）Q. Liu 和 D. Wang。Stein 变分梯度下降：一种通用贝叶斯推断算法。载于 D. Lee、M. Sugiyama、U. Luxburg、I. Guyon 和 R. Garnett 编，《神经信息处理系统进展》，第 29 卷。Curran Associates, Inc.，2016。（引用见第 13 页。）J. Lu、Y. Lu 和 J. Nolen。Stein 变分梯度下降的尺度极限：平均场区域。《SIAM数学分析杂志》，51(2):648–671，2019。（引用见第 13 页。）Y. Luo 和 N. G. Trillos。非凸矩阵分解是测地凸的：从黎曼视角对固定秩矩阵优化进行全局景观分析。arXiv 预印本 arXiv:2209.15130，2022。（引用见第 14 页。）L. Malag`o、L. Montrucchio 和 G. Pistone。高斯密度的 Wasserstein 黎曼几何。《信息几何》，1(2):137–179，2018。（引用见第 14 页。）T. Maunu、T. Le Gouic 和 P. Rigollet。Bures–Wasserstein 重心与低秩矩阵恢复。arXiv 预印本 arXiv:2210.14671，2022。（引用见第 14 页。）K. Modin。通过最优传输与信息几何视角看矩阵分解的几何学。《几何与力学杂志》，9(3):335–390，2017。（引用见第 14 页。）E. Naldi 和 G. Savar´e。Wasserstein 空间中的弱拓扑与 Opial 性质，以及其在测地凸泛函的梯度流和近端点算法中的应用。《意大利林琴国家科学院院刊：林琴数学与应用》，32(4):725–750，2021。（引用见第 8 和 30 页。）I. Olkin 和 F. Pukelsheim。具有给定离散矩阵的两个随机向量之间的距离。《线性代数及其应用》，48:257–263，1982。（引用见第 4 页。）M. Opper 和 C. Archambeau。重新审视变分高斯近似。《神经计算》，21 (3):786–792，2009。（引用见第 1 页。）F. Otto。耗散演化方程的几何学：多孔介质方程。《偏微分方程通讯》，26(1-2):101–174，2001。（引用见第 20 页。）

J. W. Paisley、D. M. Blei 和 M. I. Jordan。随机搜索的变分贝叶斯推断。载于第 29 届国际机器学习会议（ICML 2012），英国苏格兰爱丁堡，2012 年 6 月 26 日至 7 月 1 日。Omnipress，2012 年。（见第 13 页。）Z. Peng、W.-H. Wu、J. Hu 和 K.-K. Deng。流形上非光滑优化问题的黎曼平滑梯度型算法。arXiv 预印本 arXiv:2212.03526，2022 年。（见第 14 页。）G. Pleiss、M. Jankowiak、D. Eriksson、A. Damle 和 J. Gardner。快速矩阵平方根及其在高斯分布过程和贝叶斯优化中的应用。神经信息处理系统进展，33：22268–22281，2020 年。（见第 6 页。）M. Quiroz、D. J. Nott 和 R. Kohn。高维状态空间模型的高斯分布变分近似。Bayesian Analysis，1(1)：1–28，2022 年。（见第 1 页。）R. Ranganath、S. Gerrish 和 D. Blei。黑盒变分推断。载于 Artificial Intelligence and Statistics，第 814–822 页。PMLR，2014 年。（见第 13 页。）D. Rezende 和 S. Mohamed。使用归一化流的变分推断。载于国际机器学习会议，第 1530–1538 页。PMLR，2015 年。（见第 13 页。）A. Salim、A. Korba 和 G. Luise。Wasserstein 近端梯度算法。载于 H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin（编），神经信息处理系统进展，第 33 卷，第 12356–12366 页。Curran Associates, Inc.，2020 年。（见第 2、5 和 13 页。）A. Salim、L. Sun 和 P. Richtarik。在 Talagrand 不等式 T1 下总体极限中 SVGD 的收敛理论。载于国际机器学习会议，第 19139–19152 页。PMLR，2022 年。（见第 13 页。）F. Santambrogio。应用数学家的最优传输，《非线性微分方程及其应用进展》第 87 卷。Birkh¨auser/Springer，Cham，2015 年。变分法、偏微分方程与建模。（见第 20 页。）M. Seeger。支持向量机、高斯分布过程及其他核分类器的贝叶斯模型选择。神经信息处理系统进展，12，1999 年。（见第 1 页。）J. Shi 和 L. Mackey。Stein变分梯度下降的有限粒子收敛速率。arXiv 预印本 arXiv:2211.09721，2022 年。（见第 13 页。） Y. Song 、 N. Sebe 和 W. Wang。快速可微矩阵平方根和逆平方根。IEEE Transactions on Pattern Analysis and Machine Intelligence，2022 年。（见第 6 页。）V. Spokoiny。关于高维拉普拉斯近似精度的无维度非渐近界。arXiv 预印本 arXiv:2204.11038，2022 年。（见第 1 页。）P. Toulis、D. Tran 和 E. Airoldi。迈向随机梯度下降的稳定性与最优性。载于 Artificial Intelligence and Statistics，第 1290–1298 页。PMLR，2016 年。（见第 10 页。）P. Toulis、T. Horel 和 E. M. Airoldi。近端 Robbins–Monro 方法。Journal of the Royal Statistical Society Series B: Statistical Methodology，83(1)：188–212，2021 年。（见第 10 页。）A. W. Van der Vaart。渐近统计，第 3 卷。剑桥大学出版社，2000 年。（见第 1 页。）C. Villani。最优传输专题，《数学研究生教程》第 58 卷。美国数学学会，普罗维登斯，罗德岛州，2003 年。（见第 23 和 25 页。）Y. Wang、P. Chen 和 W. Li。用于高维贝叶斯推断的投影 Wasserstein 梯度下降。SIAM/ASA Journal on Uncertainty Quantification，10(4)：1513–1532，2022a。（见第 13 页。）Z. Wang、B. Liu、S. Chen、S. Ma、L. Xue 和 H. Zhao。用于稀疏谱聚类的流形近端线性方法及其在单细胞 RNA 测序数据分析中的应用。INFORMS Journal on Optimization，4(2)：200–214，2022b。（见第 14 页。）

A. Wibisono。测度空间中的采样即优化：作为复合优化问题的 Langevin 动力学。载于 S. Bubeck、V. Perchet 和 P. Rigollet（编），第 *31* 届学习理论会议论文集，第 75 卷机器学习研究论文集，第 2093–3027 页。PMLR，2018 年 7 月 6–9 日。（见第2、5、13和23页。）R. Yao 和 Y. Yang。通过 Wasserstein 梯度流进行均值场变分推断。*arXiv* 预印本 *arXiv:2207.08074*，2022 年。（见第13页。）C. Zhang、X. Chen 和 S. Ma。用于子流形上非 Lipschitz 优化的黎曼平滑最速下降方法。*arXiv* 预印本 *arXiv:2104.04199*，2021 年。（见第14页。）D. Zhang 和 S. Davanloo Tajbakhsh。用于子流形优化的黎曼随机方差缩减三次正则化牛顿方法。*Journal of Optimization Theory and Applications*，第 1–38 页，2022 年。（见第14页。）G. Zhang、S. Sun、D. Duvenaud 和 R. Grosse。作为变分推断的噪声自然梯度。载于国际机器学习会议，第 5852–5861 页。PMLR，2018 年。（见第 13 页。）

A BW空间上的微分演算

在本节中，我们推导一般泛函的 BW 梯度公式，从而为定义前向-后向算法的更新步骤提供必要工具。为此，我们给出沿测度曲线 $(\mu_t)_{t\geq 0} \subseteq \mathsf{BW}(\mathbb{R}^d)$ 对泛函 $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ 求导的计算规则，这些规则将有助于我们之后证明凸性和光滑性不等式。我们的推导基于 Otto 演算的特化；Otto 演算研究 Wasserstein 空间 $\mathcal{P}_2(\mathbb{R}^d)$，而我们将其特化到 BW 空间 $\mathsf{BW}(\mathbb{R}^d)$。

A.1 Otto演算背景

首先，我们非正式地概述 Otto 演算的计算规则（Otto，2001），该演算为 Wasserstein 空间 $\mathcal{P}_2(\mathbb{R}^d)$ 赋予了形式上的黎曼结构。关于数学理论更严谨的阐述，请参阅 Ambrosio等人 (2008)。

设 μ 是 $\mathcal{P}_2(\mathbb{R}^d)$ 中任意一个相对于勒贝格测度存在密度的元素。切空间$\mathfrak{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ 被识别为定义在 $\mathbb{R}^d$ 上的标量函数梯度所构成的空间，即

$$\mathfrak{T}_\mu\mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\psi \mid \psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}.$$

对于泛函 $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$，我们可以形式化地将其在 μ 处的 W_2 梯度定义为映射 $\nabla_{W_2}\mathcal{F}(\mu) \in \mathfrak{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$，满足

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu,$$

对于任意足够正则的测度曲线$(\mu_t)_{t\in\mathbb{R}} \subseteq \mathcal{P}_2(\mathbb{R}^d)$，满足 $\mu_0 = \mu$，以及速度向量场 $(v_t)_{t\in\mathbb{R}}$，满足 $v_t \in L^2(\mu_t)$，对几乎处处的 t 满足连续性方程

$$\partial_t\mu_t + \operatorname{div}(\mu_t v_t) = 0. \tag{26}$$

事实上，我们可以计算这个 W_2 梯度，直接通过识别计算。令 $\delta\mathcal{F}(\mu)\colon \mathbb{R}^d \to \mathbb{R}$ 表示一个一阶变分的 $\mathcal{F}$ 在 μ（参见Santambrogio，2015，第 7 章），对于该变分

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\,\partial_t|_{t=0}\mu_t.$$

然后，根据公式（26）和分部积分，

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\,\partial_t|_{t=0}\mu_t = -\int \delta\mathcal{F}(\mu)\operatorname{div}(\mu v_0) = \int \langle\nabla\delta\mathcal{F}(\mu), v_0\rangle\,\mathrm{d}\mu = \langle\nabla\delta\mathcal{F}(\mu), v_0\rangle_\mu.$$

因此，我们得出结论：

$$\nabla_{W_2}\mathcal{F}(\mu) \equiv \nabla\delta\mathcal{F}(\mu). \tag{27}$$

现在我们将注意力转向 BW空间。BW空间$\mathsf{BW}(\mathbb{R}^d)$ 是 $\mathcal{P}_2(\mathbb{R}^d)$ 的子流形（Otto，2001；Lambert 等人，2022b），因此继承了上述形式黎曼结构。

设 μ 是 $\mathsf{BW}(\mathbb{R}^d)$ 中任意一个元素。切空间 $\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ 可识别为定义在 $\mathbb{R}^d$ 上且具有对称线性项的仿射函数空间，即，

$$\mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d) = \{x \mapsto b + S\,(x - m_\mu) \mid b \in \mathbb{R}^d,\, S \in \mathbf{S}^d\}.$$

类似地，对于泛函 $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$，我们可以形式化地将其在 μ 处的 BW 梯度定义为满足以下条件的元素 $\nabla_{\mathsf{BW}}\mathcal{F}(\mu) \in \mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu), v_0\rangle_\mu,$$

对于任意测度曲线 $(\mu_t)_{t\in\mathbb{R}} \subseteq \mathsf{BW}(\mathbb{R}^d)$ 以及 $\mu_0 = \mu$ 速度向量场 $(v_t)_{t\in\mathbb{R}}$，其中每个 v_t 都是一个仿射映射，满足公式（26）。利用公式（27）和分部积分，我们将在以下小节中计算 $\mathcal{F}$ 的 BW 梯度。

A.2 BW 梯度计算

泛函的 BW 梯度 $\mathcal{F}\colon \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ 可类似地根据 Lambert 等人 (2022b，C.1 节）推导得到。为完整起见，我们在此给出推导；在此过程中，我们得到沿高斯分布曲线的 $\mathcal{F}$ 变化率公式，其中相应的速度向量场是线性部分不一定对称的仿射映射；这将在后续证明中发挥作用。关键思想是反复使用分部积分，利用高斯密度的梯度恰好是该密度乘以一个仿射项这一事实。

引理 A.1. 设 $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ 是 *Wasserstein* 空间上的一个泛函，其一阶变分为 $\delta\mathcal{F}$。那么，对于 $\mu \in \mathsf{BW}(\mathbb{R}^d)$，我们有 $\nabla_{\mathsf{BW}}\mathcal{F}(\mu)$ 由下式给出

$$\nabla_{\mathsf{BW}}\mathcal{F}(\mu)\colon x \mapsto (\mathbb{E}_\mu \nabla^2\delta\mathcal{F})(x - m_\mu) + \mathbb{E}_\mu \nabla\delta\mathcal{F}\,.$$

证明。 设 $(\mu_t)_{t\in\mathbb{R}} \subseteq \mathsf{BW}(\mathbb{R}^d)$ 是一条高斯分布正则曲线，满足 $\mu_0 = \mu$，且 $(v_t)_{t\in\mathbb{R}}$ 是一族满足方程（26）的仿射映射。此外，假设 v_0 由下式给出

$$v_0\colon x \mapsto a + M\,(x - m_\mu)\,, \qquad (a, M) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}\,,$$

并且 $\nabla_{\mathsf{BW}}\mathcal{F}(\mu) \in \mathfrak{T}_\mu\mathsf{BW}(\mathbb{R}^d)$ 由下式给出

$$\nabla_{\mathsf{BW}}\mathcal{F}(\mu)\colon x \mapsto b_\mathcal{F} + S_\mathcal{F}\,(x - m_\mu)\,, \qquad (b_\mathcal{F}, S_\mathcal{F}) \in \mathbb{R}^d \times \mathbf{S}^d\,.$$

令 $X \sim \mu$，我们得到

$$\begin{aligned}
\langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu), v_0\rangle_\mu &= \mathbb{E}\langle b_\mathcal{F} + S_\mathcal{F}\,(X - m_\mu), a + M\,(X - m_\mu)\rangle \\
&= \langle b_\mathcal{F}, a\rangle + \mathbb{E}\langle S_\mathcal{F}\,(X - m_\mu), M\,(X - m_\mu)\rangle \\
&= \langle b_\mathcal{F}, a\rangle + \mathbb{E}\langle S_\mathcal{F}, M\,(X - m_\mu)(X - m_\mu)^\mathsf{T}\rangle \\
&= \langle b_\mathcal{F}, a\rangle + \langle S_\mathcal{F}, M\Sigma_\mu\rangle \\
&= \langle b_\mathcal{F}, a\rangle + \langle S_\mathcal{F}, \Sigma_\mu M^\mathsf{T}\rangle\,. && (\text{since } S_\mathcal{F} = S_\mathcal{F}^\mathsf{T} \text{ and } \langle A, B\rangle = \langle A^\mathsf{T}, B^\mathsf{T}\rangle)
\end{aligned}$$

另一方面，根据 W_2 梯度的定义，我们得到

$$\begin{aligned}
\partial_t|_{t=0}\mathcal{F}(\mu_t) &= \langle \nabla_{W_2}\mathcal{F}(\mu), v_0\rangle_\mu && (\text{definition of } \nabla_{W_2}\mathcal{F}) \\
&= \langle \nabla\delta\mathcal{F}(\mu), v_0\rangle_\mu && (\text{by Equation (27)}) \\
&= \mathbb{E}\langle \nabla\delta\mathcal{F}(X), a + M\,(X - \mathbb{E}X)\rangle \\
&= \mathbb{E}\langle \nabla\delta\mathcal{F}(X), a\rangle + \mathbb{E}\langle \Sigma_\mu M^\mathsf{T}\nabla\delta\mathcal{F}(X), \Sigma_\mu^{-1}\,(X - \mathbb{E}X)\rangle \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle - \int \langle \Sigma_\mu M^\mathsf{T}\nabla\delta\mathcal{F}, \nabla\mu\rangle && (\text{since } \nabla\mu(x) = -\mu(x)\,\Sigma_\mu\,(x - \mathbb{E}X)) \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle + \mathbb{E}[\mathrm{div}(\Sigma_\mu M^\mathsf{T}\nabla\delta\mathcal{F})(X)] && (\text{integration by parts}) \\
&= \langle \mathbb{E}\nabla\delta\mathcal{F}(X), a\rangle + \langle \mathbb{E}_\mu\nabla^2\delta\mathcal{F}(X), \Sigma_\mu M^\mathsf{T}\rangle\,.
\end{aligned}$$

因此，通过直接识别，我们得出

$$(b_\mathcal{F}, S_\mathcal{F}) = (\mathbb{E}_\mu\nabla\delta\mathcal{F},\, \mathbb{E}_\mu\nabla^2\delta\mathcal{F})\,,$$

从而证明了我们所需的结果。 □

A.3 Ex问题的 BW 梯度和驻定条件示例(1)

考虑泛函 $\mathcal{F} = \mathcal{V} + \mathcal{H}$，其定义为势函数（与函数 V 相关）和熵之和，并回顾问题（1）等价于在 $\mathcal{F}$ BW空间上最小化，即，也就是求解

问题（19）。根据 Lambert 等人 （2022b，C.1 节），我们有以下关于 $\mathcal{V}$ 和 $\mathcal{H}$的 BW 梯度公式。

$$\begin{aligned}\nabla_{\mathsf{BW}}\mathcal{V}(\mu) &: x \mapsto \mathbb{E}_\mu \nabla V + (\mathbb{E}_\mu \nabla^2 V)(x - m_\mu)\,,\\ \nabla_{\mathsf{BW}}\mathcal{H}(\mu) &: x \mapsto -\Sigma_\mu^{-1}(x - m_\mu)\,.\end{aligned} \tag{28}$$

我们也可以从引理 A.1。

此外，根据引理 A.1的证明，我们可以计算 $\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu_t)\ v_t\rangle_{\mu_t}$ 沿任意高斯分布曲线 $(\mu_t)_{t\in\mathbb{R}}$ 以及任意仿射映射族 $(v_t)_{t\in\mathbb{R}}$，它们共同满足连续性方程。

特别地，如果 $\hat{}\pi$ 是（1）的极小化器，一阶驻点条件 $\nabla_{\mathsf{BW}}\mathcal{F}(\hat{}\pi) = 0$ 问题（1）读作

$$\mathbb{E}_{\hat{\pi}}\nabla V = 0 \qquad \text{and} \qquad \mathbb{E}_{\hat{\pi}}\nabla^2 V = \hat{\Sigma}^{-1}\,, \tag{29}$$

$\hat{}$其中 Σ 是与分布$\hat{}\pi$对应的协方差矩阵。

B 势函数和熵在 BW空间中的凸性与光滑性不等式

推导出一般泛函的BW梯度公式 $\mathcal{F}\colon\ \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ 在附录 A 中，我们现在可以继续证明引理 3.1 （关于势函数）和引理 3.2 （关于熵）。

对于这两个引理，关键思路是对一个泛函求导 $\mathcal{F}\colon\ \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ 沿着一条曲线$(\mu_t)_{t\in[0,1]}$其速度向量场 $(v_t)_{t\in[0,1]}$ 满足连续性方程(26)，利用附录中列出的计算规则A。特别地，我们将使用

$$\begin{aligned}\mathcal{F}(\mu_1) - \mathcal{F}(\mu_0) &= \int_0^1 \partial_t \mathcal{F}(\mu_t)\,\mathrm{d}t\\ &= \partial_t|_{t=0}\mathcal{F}(\mu_t) + \int_0^1\int_0^t \partial_s^2 F(\mu_s)\,\mathrm{d}s\,\mathrm{d}t\\ &= \langle \nabla_{\mathsf{BW}}\mathcal{F}(\mu_0), v_0\rangle_{\mu_0} + \int_0^1 (1-t)\,\partial_t^2\mathcal{F}(\mu_t)\,\mathrm{d}t\,,\end{aligned} \tag{30}$$

对于熵和势函数均如此。

B.1 引理的证明3.1

我们针对势函数证明以下结果。该结果强于引理3.1，并将在后续分析中发挥作用。

引理 B.1. 假设 $\alpha I \preceq \nabla^2 V \preceq \beta I$。令 $\mu \in \mathsf{BW}(\mathbb{R}^d)$，并令 $h\colon\ \mathbb{R}^d \to \mathbb{R}^d$ 为一个仿射映射。则以下不等式成立：

$$\begin{aligned}\mathcal{V}((\mathrm{id}+h)_\#\mu) - \mathcal{V}(\mu) &\geq \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_\mu + \frac{\alpha}{2}\|h\|_\mu^2\,,\\ \mathcal{V}((\mathrm{id}+h)_\#\mu) - \mathcal{V}(\mu) &\leq \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_\mu + \frac{\beta}{2}\|h\|_\mu^2\,.\end{aligned}$$

证明。 令 $X \sim \mu$。注意，无论 μ 如何，我们都有 $\delta\mathcal{V}(\mu) = V$。因此， $\nabla_{W_2}\mathcal{V}(\mu) = \nabla V$。于是我们计算得到

$$\begin{aligned}\mathcal{V}((\mathrm{id}+h)_\#\mu) - \mathcal{V}(\mu) &= \mathbb{E}[V(X+h(X)) - V(X)]\\ &\geq \mathbb{E}\big[\langle \nabla V(X), h(X)\rangle + \frac{\alpha}{2}\|h(X)\|^2\big] && (\text{since } \nabla^2 V \succeq \alpha I)\\ &= \langle \nabla_{\mathsf{W}_2}\mathcal{V}(\mu), h\rangle_\mu + \frac{\alpha}{2}\|h\|_\mu^2\\ &= \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_\mu + \frac{\alpha}{2}\|h\|_\mu^2\,,\end{aligned}$$

证明该 第一个不等式。第二个不等式的证明类似，利用以下事实： $\nabla^2 V \preceq \beta I$. □

引理3.1 因此可作为上述引理的推论得出。

引理 *3.1* 的证明。 注意，如果 V 是 β-光滑的，那么根据定义，我们有 $-\beta I \preceq \nabla^2 V \preceq \beta I$。因此，应用引理 B.1 并取 $\alpha = -\beta$，可得

$$\left|\mathcal{V}((\mathrm{id}+h)_{\#}\mu) - \mathcal{V}(\mu) - \langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu), h\rangle_{\mu}\right| \le \frac{\beta}{2}\|h\|_{\mu}^2.$$

此外，我们在附录 A.3 中已证明 $\nabla_{\mathsf{BW}}\mathcal{V}(\mu)$ 的表达式为

$$\nabla_{\mathsf{BW}}\mathcal{V}(\mu) : x \mapsto \mathbb{E}_{\mu}\nabla V + (\mathbb{E}_{\mu}\nabla^2 V)(x - m_{\mu}),$$

至此，我们所需结果的证明完成。 □

B.2 引理3.2的证明

对于熵，我们沿用前一个证明中的相同策略，沿着一条特定曲线对熵 $\mathcal{H}$ 求导。这次，我们将沿着如下定义的广义测地线 $(\mu_t^{\nu})_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$求导：

设 T_0, T_1 为最优传输映射，使得 $T_0 - \mathrm{id}$ $T_1 - \mathrm{id} \in T_{\nu}\mathsf{BW}(\mathbb{R}^d)$，且分别满足 $(T_0)_{\#}\nu = \mu_0$ 和 $(T_1)_{\#}\nu = \mu_1$。定义 $T_t := (1-t)T_0 + tT_1$，以 ν 为基点、μ_0, μ_1 为端点的广义测地线，便是测度曲线 $(\mu_t^{\nu})_{t\in[0,1]} \subseteq \mathsf{BW}(\mathbb{R}^d)$，其中 $\mu_t^{\nu} = (T_t)_{\#}\nu$。注意，$\mu_0^{\nu} = \mu_0$和 $\mu_1^{\nu} = \mu_1$，且 $(\mu_t^{\nu})_{t\in[0,1]}$ 满足连续性方程

$$\partial_t \mu_t^{\nu} + \mathrm{div}(\mu_t^{\nu} v_t) = 0, \qquad \text{where } v_t = (T_1 - T_0)\circ T_t^{-1}.$$

广义测地线被用于 Ambrosio等人 (2008) 研究 Wasserstein 空间中的梯度流，此后也被用于该理论的各种应用，例如 Chewi等人 (2020b); Ahn 和 Chewi(2021); Altschuler等人(2021)。

引理 *3.2* 的证明。 在 Wasserstein 空间 $\mathcal{H}$ 上作用于 $\mathcal{P}_2(\mathbb{R}^d)$ 的 JKO 算子由 Wibisono(2018，例 7）针对高斯测度 $\mu = \mathcal{N}(\mu,\ \Sigma)$ 推导得出，其形式为 $\mu' = \mathcal{N}(\mu,\ \Sigma_1)$，其中 Σ_1 的定义方式与方程（17）相同。由于 μ' 也是 $\mathsf{BW}(\mathbb{R}^d)$ 的一个元素，我们得出 μ' 也是将 BW JKO 算子作用于 μ 所得的结果，从而证明了我们所需的闭式表达。

现在我们证明所需的熵的广义测地凸性不等式。事实上，这一结论可由 Wasserstein 空间上的一般结果推出（例如，见 Ambrosio等人，2008，§9.4），但为完整起见，我们在此给出证明。如上所述，为此我们将沿着上述定义的广义测地线（$\mathcal{H}$ $(\mu_t^{\nu})_{t\in[0,1]}$）对其求导。为方便起见，我们将分布 μ 与其相对于勒贝格测度的密度等同起来。于是有

$$\begin{aligned}
\partial_t^2 \mathcal{H}(\mu_t^{\nu}) &= \partial_t^2 \int \mu_t^{\nu}\ln\mu_t^{\nu} = \partial_t^2 \int \nu \ln(\mu_t^{\nu}\circ T_t) = \partial_t^2 \int \nu \ln\frac{\nu}{\det\nabla T_t} \quad (\text{since } (T_t)_{\#}\nu = \mu_t^{\nu},\ \text{change of variable})\\
&= -\int (\partial_t^2 \ln\det\nabla T_t)\,\mathrm{d}\nu = -\int \partial_t\left\langle [\nabla T_t]^{-1}, \partial_t \nabla T_t\right\rangle \mathrm{d}\nu\\
&= -\int \partial_t\left\langle [\nabla T_t]^{-1}, \nabla T_1 - \nabla T_0\right\rangle \mathrm{d}\nu = \int \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2\right\rangle \mathrm{d}\nu\\
&= \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2\right\rangle,
\end{aligned}$$

其中最后一行成立，是因为 T_t 是一个仿射映射，这意味着 ∇T_t 在 $\mathbb{R}^d$上为常数。此外，根据布雷尼尔定理（Villani，2003，定理 2.12），T_t 对所有 $t \in [0,1]$都是某个凸函数的梯度。因此，我们知道，对所有 $\nabla T_t \succeq 0$，$t \in [0,1]$，这意味着 $\left\langle [\nabla T_t]^{-2},\ (\nabla T_1 - \nabla T_0)^2\right\rangle \ge 0$. 因此，将方程（30）应用于 $\mathcal{H}$，可得

$$\begin{aligned}
\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)\circ T_0, T_1 - T_0\rangle_{\nu} + \int_0^1 (1-t)\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2\right\rangle \mathrm{d}t\\
&\ge \langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)\circ T_0, T_1 - T_0\rangle_{\nu}.
\end{aligned}$$

这就证明了熵所满足的所需不等式，我们的证明也随之完成。 □

备注 B.2。 事实上，我们可以证明熵沿着连接具有相同均值的分布 μ_0， $\mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ 的广义测地线满足一个强凸性 不等式。设 m_0， m_1 分别为 μ_0， μ_1 的均值，并假设 Σ_{μ_0}， $\Sigma_{\mu_1} \preceq \lambda I$。我们计算得

$$\begin{aligned}
\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle &= \left\langle I, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\
&\geq \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle \Sigma_{\mu_t}^\nu, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\
&= \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle [\nabla T_t]^{-1} \Sigma_{\mu_t^\nu} [\nabla T_t]^{-1}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\
&= \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle.
\end{aligned}$$

由于 T_0 是仿射映射，我们知道，对于所有 $x \in \mathbb{R}^d$， $T_0(x) - (\nabla T_0)\, x$ 是一个常数， T_1也同样如此。因此，如果 $Y \sim \nu$，则

$$\begin{aligned}
\|T_1 - T_0\|_\nu^2 &= \mathrm{Tr}(\mathrm{Cov}_\nu[T_1 - T_0, T_1 - T_0]) + \|\mathbb{E}_\nu[T_1 - T_0]\|^2 && \text{(by bias-variance decomposition)} \\
&= \mathrm{Tr}(\mathrm{Cov}[(\nabla T_1 - \nabla T_0)(Y), (\nabla T_1 - \nabla T_0)(Y)]) + \|m_1 - m_0\|^2 && \text{(since } T_0, T_1 \text{ are affine)} \\
&= \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle + \|m_1 - m_0\|^2 . && (31)
\end{aligned}$$

此外，根据Chewi 等人（2020b，引理 10），我们知道，协方差矩阵的算子范数沿着BW（$\mathbb{R}^d$）中的广义测地线是凸的，这意味着 $\Sigma_{\mu_t^\nu} \preceq \lambda I$ 对所有 $t \in [0,1]$均成立。因此，我们得到

$$\begin{aligned}
\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle &\leq \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\
&= \frac{1}{\left\|\Sigma_{\mu_t^\nu}\right\|_{\mathrm{op}}} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right) && \text{(by Equation (31))} \\
&\geq \frac{1}{\lambda} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right). && \text{(by Chewi et al. (2020b, Lemma 10))}
\end{aligned}$$

因此，将等式（30）应用于 $\mathcal{H}$，可得

$$\begin{aligned}
\mathcal{H}(\mu_1) - \mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \int_0^1 (1-t) \left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \mathrm{d}t \\
&\geq \langle \nabla_{\mathsf{BW}} \mathcal{H}(\mu_0) \circ T_0, T_1 - T_0 \rangle_\nu + \frac{1}{2\lambda} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right).
\end{aligned}$$

这意味着，对于均值相同的两个高斯分布 μ_0， $\mu_1 \in \mathsf{BW}(\mathbb{R}^d)$，熵沿广义测地线呈强凸性。类似地，通过相同的计算可以证明，熵沿测地线满足光滑性 不等式。和之前一样，我们计算得到

$$\begin{aligned}
\left\langle [\nabla T_t]^{-2}, (\nabla T_1 - \nabla T_0)^2 \right\rangle &= \left\langle I, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\
&\leq \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left\langle \Sigma_{\mu_t}^\nu, [\nabla T_t]^{-1} (\nabla T_1 - \nabla T_0)^2 [\nabla T_t]^{-1} \right\rangle \\
&= \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left\langle [\nabla T_t]^{-1} \Sigma_{\mu_t^\nu} [\nabla T_t]^{-1}, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\
&= \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left\langle \Sigma_\nu, (\nabla T_1 - \nabla T_0)^2 \right\rangle \\
&= \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \left(\|T_1 - T_0\|_\nu^2 - \|m_1 - m_0\|^2 \right) \\
&\leq \frac{1}{\lambda_{\min}(\Sigma_{\mu_t^\nu})} \|T_1 - T_0\|_\nu^2 .
\end{aligned}$$

再次将方程（30）应用于 $\mathcal{H}$，可得

$$\begin{aligned}\mathcal{H}(\mu_1)-\mathcal{H}(\mu_0) &= \langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)\circ T_0, T_1-T_0\rangle_\nu + \int_0^1 (1-t)\left\langle [\nabla T_t]^{-2}, (\nabla T_1-\nabla T_0)^2\right\rangle \mathrm{d}t \\ &\le \langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)\circ T_0, T_1-T_0\rangle_\nu + \int_0^1 \frac{1-t}{\lambda_{\min}(\Sigma_{\mu_t^\nu})}\|T_1-T_0\|_\nu^2\,\mathrm{d}t\,. \end{aligned} \tag{32}$$

作为不等式 32的推论，我们得到了熵沿测地线的光滑性不等式，这将对后续分析有所帮助。

引理 B.3 (熵沿测地线的光滑性)。 假设 $\mu_0,\mu_1\in\mathsf{BW}(\mathbb{R}^d)$ 满足 $\Sigma_{\mu_0}^{-1},\Sigma_{\mu_1}^{-1}\preceq\gamma I$。那么，如果 T 是从 μ_0 到 μ_1 的最优传输映射，则有

$$\mathcal{H}(\mu_1)-\mathcal{H}(\mu_0)\le\langle\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0), T-\mathrm{id}\rangle_{\mu_0}+\frac{\gamma}{2}\|T-\mathrm{id}\|_{\mu_0}^2\,.$$

证明。 我们对 32 应用不等式， $\nu=\mu_0$注意到在此情形下 $T_1=T$ 且 $T_0=$ 为恒等映射，并且 $(\mu_t^\nu)_{t\in[0,1]}$恰好是连接 $\mu_t)_{t\in[0,1]}$ 和 μ_0,μ_1的等速测地线。此外，根据 Altschuler 等人 (2021，附录 B)，我们知道 $\lambda_{\min}$ 沿测地线是凹的，因此 $\lambda_{\min}(\Sigma_{\mu_t})\ge\gamma^{-1}I$ 对所有 t均成立。因此，我们得到

$$\begin{aligned}\mathcal{H}(\mu_1)-\mathcal{H}(\mu_0) &\le \langle\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0),T-\mathrm{id}\rangle_{\mu_0}+\int_0^1\frac{1-t}{\lambda_{\min}(\Sigma_{\mu_t})}\|T-\mathrm{id}\|_{\mu_0}^2\,\mathrm{d}t \\ &\le \langle\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0),T-\mathrm{id}\rangle_{\mu_0}+\int_0^1\frac{1-t}{\gamma^{-1}}\|T-\mathrm{id}\|_{\mu_0}^2\,\mathrm{d}t \\ &= \langle\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0),T-\mathrm{id}\rangle_{\mu_0}+\frac{\gamma}{2}\|T-\mathrm{id}\|_{\mu_0}^2\,,\end{aligned}$$

从而证明所需结果。 □

C 一步不等式的证明（引理5.1）

该证明的 关键思路是将差值 $\mathcal{F}(p_{k+1})-\mathcal{F}(\nu)$ 分解为三项之和，

$$\mathcal{F}(p_{k+1})-\mathcal{F}(\nu)=[\mathcal{V}(p_{k+1})-\mathcal{V}(p_{k+\frac12})]+[\mathcal{V}(p_{k+\frac12})-\mathcal{V}(\nu)]+[\mathcal{H}(p_{k+1})-\mathcal{H}(\nu)]\,,$$

其中，每一项都可以利用引理 3.2 和 B.1中的不等式进行控制。回顾引理 3.2 仅适用于广义测地线，因此我们必须谨慎地定义p_k、$p_{k+\frac12}$、p_{k+1}和 ν之间的耦合。以下证明将详细说明这一论证。

引理 *5.1* 的证明。回顾第 5 节中的定义， $\mathscr{F}_k$ 是由 σ生成的、迭代至 k为止的代数（但不包括随机样本 $X_k\sim p_k$，该样本来自随机 FB–GVI）。此外，我们还有

$$e_k\colon x\mapsto(S_k-\mathbb{E}_{p_k}\nabla^2 V)(x-m_k)+(b_k-\mathbb{E}_{p_k}\nabla V)$$

定义为（随机）FB–GVI（随机）梯度估计在第 k 次迭代时的误差，其中 $\mathbb{E}[e_k\mid\mathscr{F}_k]=0$。在给定滤过 $\mathscr{F}_k$ 的条件下，我们构造以下随机变量X_k， $X_{k+\frac12}$， X_{k+1}， $Y_\mathcal{V}$和 $Y_\mathcal{H}$。

令 $(X_k,\ Y_\mathcal{V})\sim(p_k,\ \nu)$ 在 W_2 距离下构成最优耦合，并令 $(X_k,\ Y_\mathcal{V})\perp\!\!\!\perp e_k$。根据假设，由于 $\eta\le\frac{1}{\beta}$，我们有–

$$I-\eta S_k\succeq(1-\eta\beta)\,I\succeq 0\,.$$

回顾布雷尼尔定理（Villani，2003，定理 2.12）：若 $Y=\nabla\varphi(X$是一个凸、适当且下半连续函数 $\varphi\colon\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$，则（$X$，$Y$）是 2-Wasserstein 距离下的最优耦合。因此，上述条件 $I-\eta S_k\succeq 0$保证了（X_k，$X_{k+\frac12}$）$\sim(p_k,\ p_{k+\frac12})$ 是 W_2距离下的最优耦合，其中我们定义

$$X_{k+\frac12}:=(I-\eta S_k)(X_k-m_k)+m_k-\eta b_k\,.$$

另一方面，定义 X_{k+1} 使得

$$\begin{aligned} X_{k+1} &\coloneqq X_{k+\frac{1}{2}} - \eta\, \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}] \\ &= (I + \eta\, \Sigma_{k+1})^{-1} (X_{k+\frac{1}{2}} - m_{k+1}) + m_{k+1}\,, \end{aligned}$$

我们还得到，$(X_{k+\frac{1}{2}},\ X_{k+1}) \sim (p_{k+\frac{1}{2}},\ p_{k+1})$ 是最优耦合的。最后，我们构造随机变量 $Y_{\mathcal{H}} \sim \nu$，使得 $(X_{k+\frac{1}{2}},\ Y_{\mathcal{H}})$ 关于 W_2 距离是最优耦合的。

首先，我们对能量差进行界定。根据布雷尼尔定理，我们知道 $Y_{\mathcal{H}}$ 和 X_{k+1} 都可以表示为 X_k的仿射函数，因此可以应用引理 B.1。由此可得

$$\begin{aligned} \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(\nu)] &= \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(p_k)] + \mathbb{E}[\mathcal{V}(p_k) - \mathcal{V}(\nu)] \\ &\le \mathbb{E}\langle \nabla_{\mathsf{BW}} \mathcal{V}(p_k)(X_k), X_k - Y_{\mathcal{V}} \rangle - \frac{\alpha}{2}\, \mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 \\ &\quad + \mathbb{E}\langle \nabla_{\mathsf{BW}} \mathcal{V}(p_k)(X_k), X_{k+1} - X_k \rangle + \frac{\beta}{2}\, \mathbb{E}\|X_{k+1} - X_k\|^2 \qquad \text{(by Lemma B.1)} \\ &= -\frac{\alpha}{2}\, \mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 + \mathbb{E}\langle \nabla_{\mathsf{BW}} \mathcal{V}(p_k)(X_k), X_{k+1} - Y_{\mathcal{V}} \rangle \\ &\quad + \frac{1}{2\eta}\, \mathbb{E}\|X_{k+1} - X_k\|^2 - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\, \mathbb{E}\|X_{k+1} - X_k\|^2 \\ &= -\frac{\alpha}{2}\, \mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}} \rangle - \frac{1}{\eta}\, \mathbb{E}\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}} \rangle \\ &\quad + \frac{1}{2\eta}\, \mathbb{E}\|X_{k+1} - X_k\|^2 - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\, \mathbb{E}\|X_{k+1} - X_k\|^2 \\ &= \frac{1}{2\eta}\, (1 - \alpha\eta)\, \mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}} \rangle - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\, \mathbb{E}\|X_{k+1} - X_k\|^2 \\ &\quad + \frac{1}{2\eta}\, \mathbb{E}\big[\|X_{k+1} - X_k\|^2 - \|X_k - Y_{\mathcal{V}}\|^2 - 2\,\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}} \rangle\big]\,. \end{aligned}$$

现在我们对熵差进行界定。由于 $Y_{\mathcal{H}}$和 X_{k+1}都与 $X_{k+\frac{1}{2}}$最优耦合，我们知道 $(Y_{\mathcal{H}},\ X_{k+1})$ 沿着一条广义测地线耦合。因此，我们可以应用引理 3.2 得到

$$\begin{aligned} \mathbb{E}[\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)] &\le \mathbb{E}\langle \nabla_{\mathsf{BW}} \mathcal{H}(p_{k+1})[X_{k+1}], X_{k+1} - Y_{\mathcal{H}} \rangle \\ &= -\frac{1}{\eta}\, \mathbb{E}\langle X_{k+1} - X_{k+\frac{1}{2}}, X_{k+1} - Y_{\mathcal{H}} \rangle \\ &= \frac{1}{2\eta}\, \mathbb{E}\big[\|X_{k+\frac{1}{2}} - Y_{\mathcal{H}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] \\ &\le \frac{1}{2\eta}\, \mathbb{E}\big[\|X_{k+\frac{1}{2}} - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big]\,. \\ &\qquad \text{(since } (X^{k+\frac{1}{2}}, Y_{\mathcal{H}}) \text{ are optima)} \end{aligned}$$

现在，我们将 利用上述不等式可得到我们所需的关于 $\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)]$ 的界。我们得到

$$\begin{aligned} \mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &= \mathbb{E}[\mathcal{V}(p_{k+1}) - \mathcal{V}(\nu)] + \mathbb{E}[\mathcal{H}(p_{k+1}) - \mathcal{H}(\nu)] \\ &\le \frac{1}{2\eta}\, \mathbb{E}\big[(1 - \alpha\eta)\, \|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] \\ &\quad + \frac{1}{2\eta}\, \mathbb{E}\big[\|X_{k+1} - X_k\|^2 - \|X_k - Y_{\mathcal{V}}\|^2 + \|X_{k+\frac{1}{2}} - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - X_{k+\frac{1}{2}}\|^2\big] \\ &\quad - \frac{1}{2\eta}\, \mathbb{E}\big[2\,\langle X_{k+\frac{1}{2}} - X_k, X_{k+1} - Y_{\mathcal{V}} \rangle\big] \\ &\quad - \mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}} \rangle - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\, \mathbb{E}\|X_{k+1} - X_k\|^2 \\ &= \frac{1}{2\eta}\, \mathbb{E}\big[(1 - \alpha\eta)\, \|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] \end{aligned}$$

$$-\mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\|X_{k+1} - X_k\|^2 . \tag{33}$$

最后，还需要对最后一行中的误差项进行界定。为此，我们根据误差项 e_k 是否恒等于零，分两种情况讨论：

- 在 FB–GVI 情形下，我们可以获得精确梯度 $\nabla_{\mathrm{BW}}\mathcal{V}(p_k)$，有 $e_k \equiv 0$，因此

  $$-\mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle = 0 .$$

  结合不等式 33，我们得到，在 $\eta \le \frac{1}{\beta}$ 下，

  $$\begin{aligned}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &\le \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] - \Big(\frac{1}{2\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\ &\le \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] .\end{aligned}$$

  重新整理可得，如果 $e_k \equiv 0$ 且 $\eta \le \frac{1}{\beta}$，

  $$\begin{aligned}\mathbb{E}W_2^2(p_{k+1}, \nu) &\le \mathbb{E}\|X_{k+1} - Y_{\mathcal{H}}\|^2 &(34)\\ &\le (1-\alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &(35)\\ &= (1-\alpha\eta)\,\mathbb{E}W_2^2(p_k, \nu) - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] .\end{aligned}$$

  （由于在 $_k$,di i $_k$,d $_\nu$ 是最优耦合的）

- 否则，如果 e_k 不一定恒等于 0，我们仍然可以计算

  $$\begin{aligned}-\mathbb{E}\langle e_k(X_k), X_{k+1} - Y_{\mathcal{V}}\rangle &= -\mathbb{E}\langle e_k(X_k), X_{k+1} - X_k\rangle \qquad (\text{since } e_k \perp\!\!\!\perp (X_k, Y_{\mathcal{V}}) \text{ by construction}) \\ &\le \eta\,\mathbb{E}\|e_k(X_k)\|^2 + \frac{1}{4\eta}\,\mathbb{E}\|X_{k+1} - X_k\|^2 . \\ &\qquad (\text{Cauchy–Schwarz and Young's inequality})\end{aligned}$$

  因此，结合不等式 33，我们得到对于 $\eta \le \frac{1}{2\beta}$，

  $$\begin{aligned}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] &\le \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] + \eta\,\mathbb{E}\|e_k(X_k)\|^2 \\ &\quad - \Big(\frac{1}{4\eta} - \frac{\beta}{2}\Big)\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\ &\le \frac{1}{2\eta}\,\mathbb{E}\big[(1-\alpha\eta)\,\|X_k - Y_{\mathcal{V}}\|^2 - \|X_{k+1} - Y_{\mathcal{H}}\|^2\big] + \eta\,\mathbb{E}\sigma_k^2 .\end{aligned}$$

  重新整理后，我们得出，只要 $\eta \le \frac{1}{2\beta}$，

  $$\begin{aligned}\mathbb{E}W_2^2(p_{k+1}, \nu) &\le \mathbb{E}\|X_{k+1} - Y_{\mathcal{H}}\|^2 \\ &\le (1-\alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\,\mathbb{E}\sigma_k^2 \\ &= (1-\alpha\eta)\,\mathbb{E}W_2^2(p_k, \nu) - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] + 2\eta^2\,\mathbb{E}\sigma_k^2 . \\ &\qquad (\text{since } (X_k, Y_{\mathcal{V}}) \text{ are optimally coupled})\end{aligned}$$

结合这两种情况，我们证明了所需的不等式。 □

注C.1。 考虑将上述证明专门应用于 $\nu = p_k$ 的情形，此时 $Y_{\mathcal{V}} = Y_{\mathcal{H}} = X_k$，从而 $(X_k,\ X_{k+\frac{1}{2}}) \sim (p_k,\ p_{k+\frac{1}{2}})$ 以及 $(X_{k+\frac{1}{2}},\ X_{k+1}) \sim (p_{k+\frac{1}{2}},\ p_{k+1})$ 关于 W_2 距离是最优耦合的。于是由不等式 35，我们得到

$$\begin{aligned}&\mathbb{E}\|X_{k+1} - Y_{\mathcal{H}}\|^2 \le (1-\alpha\eta)\,\mathbb{E}\|X_k - Y_{\mathcal{V}}\|^2 - 2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(\nu)] \qquad (\text{Inequality 35}) \\ \Longrightarrow\ &\mathbb{E}\|X_{k+1} - X_k\|^2 \le -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)] . \qquad (\text{since } \nu = p_k \text{ and } Y_{\mathcal{V}} = Y_{\mathcal{H}} = X_k)\end{aligned}$$

As 作为推论，我们得到以下引理，这将在后续分析中发挥作用 .

引理 *C.2.* 假设 V 是 β-光滑的。设 $(p_k)_{k\in\mathbb{N}}$ 为 *FB–GVI (20)–(21)* 的迭代序列。设 $\eta>0$ 满足 $\eta\le\frac{1}{\beta}$。设 $(X_k, X_{k+\frac{1}{2}}) \sim (p_k, p_{k+\frac{1}{2}})$ 和 $(X_{k+\frac{1}{2}}, X_{k+1}) \sim (p_{k+\frac{1}{2}}, p_k)$ 关于 W_2 距离是最优耦合的。那么，

$$\mathbb{E}\|X_{k+1}-X_k\|^2 \le -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1})-\mathcal{F}(p_k)]\,.$$

D 迭代序列的特征值控制

我们将证明以下特征值界结果：

引理 D.1。 在 k 次迭代算法 *1* 中，假设我们有 $\gamma_0 I \preceq \Sigma_k^{-1} \preceq \gamma_1 I$。只要$0\le\eta\le\frac{1}{\gamma_1}$ 并且 $\gamma_0 I \preceq S_k \preceq \gamma_1 I$，那么我们有。

$$\gamma_1^{-1} I \preceq \Sigma_{k+1} \preceq \gamma_0^{-1} I\,.$$

证明。 定义单调递增函数 $f_\eta\colon \mathbb{R}_{\ge0}\to\mathbb{R}_{\ge0}$ 使得

$$f_\eta(x)=\frac{1}{2}\left(x+2\eta+\sqrt{x\,(x+4\eta)}\right).$$

首先，我们注意以下代数恒等式。定义 $x_\gamma := (1-\eta\gamma)^2/\gamma$。则有

$$\begin{aligned} f_\eta(x_\gamma) &= \frac{1}{2}\left(\frac{(1-\eta\gamma)^2}{\gamma}+2\eta+\sqrt{\left(\frac{(1-\eta\gamma)^2}{\gamma}\right)\left(\frac{(1-\eta\gamma)^2}{\gamma}+4\eta\right)}\right)\\ &= \frac{1}{2\gamma}\left(1+\eta^2\gamma^2+\sqrt{(1-\eta\gamma)^2\,(1+\eta\gamma)^2}\right)\\ &= \frac{1}{2\gamma}\left(1+\eta^2\gamma^2+(1-\eta\gamma)\,(1+\eta\gamma)\right)\\ &= \frac{1}{\gamma}\,. \end{aligned} \tag{36}$$

现在，令 $\lambda_{\min}(M)$ $\lambda_{\max}(M)$ 表示矩阵 $M\in\mathbf{S}^d$ 的最小特征值和最大特征值。条件 $\eta\le\gamma_1^{-1}$和 $S_k\preceq\gamma_1 I$ 则意味着 $I-\eta S_k\succeq 0$。因此，我们有

$$\begin{aligned} \lambda_{\min}(\Sigma_{k+\frac{1}{2}}) &= \lambda_{\min}\big((I-\eta S_k)\,\Sigma_k\,(I-\eta S_k)\big)\\ &\ge \lambda_{\min}^2(I-\eta S_k)\,\lambda_{\min}(\Sigma_k)\\ &\ge (1-\eta\gamma_1)^2\,\lambda_{\min}(\Sigma_k)\\ &\ge \frac{(1-\eta\gamma_1)^2}{\gamma_1}\\ &= x_{\gamma_1}\,. \end{aligned}$$

现在，我们还注意到， $\Sigma_{k+\frac{1}{2}}$ 和 Σ_{k+1} 按构造可交换，因此由于 f_η 是单调递增函数，

$$\lambda_{\min}(\Sigma_{k+1}) = f_\eta\big(\lambda_{\min}(\Sigma_{k+\frac{1}{2}})\big) \ge f_\eta(x_{\gamma_1}) = \frac{1}{\gamma_1}\,,$$

其中最后一个等式由公式（36）得出。

同理，对于上界，我们有

$$\begin{aligned} \lambda_{\max}(\Sigma_{k+\frac{1}{2}}) &= \lambda_{\max}\big((I-\eta S_k)\,\Sigma_k\,(I-\eta S_k)\big)\\ &\le \lambda_{\max}^2(I-\eta S_k)\,\lambda_{\max}(\Sigma_k) && (\text{since } I-\eta S_k\succeq 0)\\ &\le (1-\eta\gamma_0)^2\,\lambda_{\max}(\Sigma_k) \end{aligned}$$

$$\leq \frac{(1-\eta\gamma_0)^2}{\gamma_0}.$$

因此，同理可得

$$\lambda_{\max}(\Sigma_{k+1}) = f_\eta\big(\lambda_{\max}(\Sigma_{k+\frac{1}{2}})\big) \leq f_\eta(x_{\gamma_0}) = \frac{1}{\gamma_0}.$$

结合上述结果，这就证明了 $\gamma_1^{-1} I \preceq \Sigma_{k+1} \preceq \gamma_0^{-1} I$，也正是我们想要证明的。 □

请注意，对于（随机）FB–GVI，我们有 $\alpha I \preceq S_k \preceq \beta I$，因此引理 D.1 在 $\gamma_0 = \alpha$ 和 $\gamma_1 = \beta$下成立。因此，我们得到以下推论：

推论 D.2。 假设算法 *1* 以矩阵 Σ_0 初始化，满足 $\beta^{-1} I \preceq \Sigma_0$，且 V 是β光滑的，并且步长满足 $\eta \leq \frac{1}{\beta}$。那么 $\beta^{-1} I \preceq \Sigma_k$ 对于所有 k。

E 无噪声算法收敛速率的证明

通过重新排列并迭代引理5.1中的一步不等式，我们获得了FB–GVI所需的收敛速率。首先，我们推导出同时适用于凸情形和强凸情形的不等式。

对于 FB–GVI，我们可以应用引理 5.1，其中 $\nu = \hat{\pi}$，$\eta \leq \frac{1}{\beta}$ 以及 $\sigma_k = 0$。此外，FB–GVI 是确定性的，因此我们可以去掉引理5.1中的期望。在这种情况下，引理 5.1中的不等式意味着，对于所有 k，

$$\begin{aligned} W_2^2(p_{k+1}, \hat{\pi}) &\leq (1-\alpha\eta)\, W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})) && \text{(by Lemma 5.1)} \\ &\leq \exp(-\alpha\eta)\, W_2^2(p_k, \hat{\pi}) - 2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}))\,. && (37) \end{aligned}$$

重新整理公式（37），我们得到

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}) \leq \frac{\exp(-\alpha\eta)\, W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})}{2\eta}. \tag{38}$$

另一方面，我们也可以应用引理 5.1，取 $\nu = p_k$，$\eta \leq \frac{1}{\beta}$ 和 $\sigma_k^2 = 0$，从而得到

$$W_2^2(p_{k+1}, p_k) \leq (1-\alpha\eta)\, W_2^2(p_k, p_k) - 2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)) = -2\eta\,(\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k))\,.$$

因此，重新整理该不等式可得，

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k) \leq -\frac{W_2^2(p_{k+1}, p_k)}{2\eta} \leq 0\,, \tag{39}$$

这意味着算法的目标值在每次迭代后都会降低。

E.1 定理的证明 5.2

证明。 由于 V是凸的，不等式 38 在取 $\alpha = 0$时成立，由此可得

$$\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi}) \leq \frac{W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})}{2\eta}.$$

对该不等式求和并消去中间项，可得

$$\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \frac{1}{N}\sum_{k=1}^{N}[\mathcal{F}(p_k) - \mathcal{F}(\hat{\pi})] \leq \frac{1}{2\eta N}\sum_{k=0}^{N-1}[W_2^2(p_k, \hat{\pi}) - W_2^2(p_{k+1}, \hat{\pi})] \leq \frac{W_2^2(p_0, \hat{\pi})}{2\eta N}\,,$$

其中，第一个不等式由不等式 39得出。因此，取

$$\eta = \frac{1}{\beta}\,, \quad \text{and} \quad N \gtrsim \frac{\beta W_2^2(p_0, \hat{\pi})}{\varepsilon^2}\,,$$

我们得到保证 $\mathcal{F}(p_N) - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2$, 从而证明了我们所期望的结果。 □

E.2 定理 5.3 的证明

证明。 我们的证明利用了Naldi 和 Savar´e （2021）的最新结果。对于每个连续函数$\zeta: \mathbb{R}^d \to \mathbb{R}$，记 F_ζ 为定义在 $\mathcal{P}_2(\mathbb{R}^d)$ 上的映射 $F_\zeta: \mu \mapsto \int \zeta \mathrm{d}\mu$。我们将$\mathcal{P}_2(\mathbb{R}^d)$ 上由函数族 F_ζ 诱导的初始拓扑定义为 NS 拓扑，其中 ζ 是满足$\frac{\zeta(x)}{1+\|x\|^2} \longrightarrow_{\|x\|\to\infty} 0.$ 的连续函数。记 $\mathcal{C}$ 为此类 ζ 函数的集合，序列 $(\mu_k)_{k\geq 0}$ 在 NS 拓扑下收敛到 $\hat{\mu}$，当且仅当对于每个 $\zeta \in \mathcal{C}$，$F_\zeta(\mu_k) \to F_\zeta(\hat{\mu})$。

有界连续函数集合在 $\mathbb{R}^d$ 中包含于 $\mathcal{C}$，因此 NS 收敛意味着弱收敛。此外，满足 ζ 上确界$\sup_{x\in\mathbb{R}^d} \frac{\zeta(x)}{1+\|x\|^2} < \infty$ 的连续函数集合包含 $\mathcal{C}$，因此 Wasserstein 距离上的收敛意味着 NS 收敛，参见 Naldi and Savar´e(2021，公式 1.9 和 1.10）。

为证明定理 5.3，我们应用Naldi and Savar´e (2021，定理 6.9）。该定理推出 FB–GVI 的 NS 收敛，而这又意味着弱收敛。现在我们检验 NaldiandSavar´e(2021，定理 6.9）的假设：

- 首先，集合 $A = \mathrm{BW}(\mathbb{R}^d) \cup \mathcal{D}$ 是 NS 闭的，因为它在弱拓扑下是闭的；其中 $\mathcal{D}$ 是 $\mathbb{R}^d$ 上的狄拉克测度集合。
- 接下来，记 $T: A \to A$ 定义 FB–GVI 的映射（见公式 (21)），即 $T(\mu) = \mathrm{JKO}_{\eta\mathcal{H}}((\mathrm{id} - \eta\nabla_{\mathrm{BW}}\mathcal{V}(\mu))_{\#}\mu)$ 和 $p_{k+1} = T(p_k)$，其中 $p_0 \in A$。我们证明 T是渐近正则的。利用引理 5.1 并结合 $\alpha = 0$，若 $\eta \leq 1/\beta$，则

$$W_2^2(T(p_k), \nu) \leq W_2^2(p_k, \nu) - 2\eta\left[\mathcal{F}(T(p_k)) - \mathcal{F}(\nu)\right]. \tag{40}$$

查看引理 5.1 的证明，尤其是公式 (34)，可以发现，当 $\eta < 1/\beta$ 时，我们实际上有以下更强的不等式：

$$W_2^2(T(p_k), \nu) \leq W_2^2(p_k, \nu) - 2\eta\left[\mathcal{F}(T(p_k)) - \mathcal{F}(\nu)\right] - (1-\eta\beta)\, W_2^2(T(p_k), p_k). \tag{41}$$

此外，若 ν是问题 (1) 的极小化器，则 $\mathcal{F}(T(p_k)) - \mathcal{F}(\nu) \geq 0$，因此

$$W_2^2(T(p_k), \nu) \leq W_2^2(p_k, \nu) - (1-\eta\beta)\, W_2^2(T(p_k), p_k). \tag{42}$$

对最后一个不等式进行迭代，可得 $\sum_{k\geq 0} W_2^2(T(p_k)\ p_k) < \infty$。特别地，对于任意初始测度 $p_0 \in A$，有 $W_2^2(T(p_k)\ p_k) \to 0$。因此， T 是渐近正则的。

- 关于非扩张性的证明 T 与渐近正则性的证明类似：可以调整引理 5.1 来证明一个一步不等式，其中 ν 在变化并遵循 FB–GVI 算法。由于这部分证明与引理 5.1的证明类似，我们略去这部分证明。
- 最后，利用公式 $(p_k)_{k\geq 0}$ 对其进行界定（42）。

因此，根据 Naldi 和Savar´e（2021，定理 6.9），$(p_k)_{k\geq 0}$ 弱收敛到某个 $\pi_\star \in A$。此外，根据定理5.2，$\mathcal{F}(p_k) - \min_{\mathrm{BW}(\mathbb{R}^d)}\mathcal{F} \to 0$。因此，利用KL散度的弱下半连续性，我们有 $\mathcal{F}(\pi_\star) - \min_{\mathrm{BW}(\mathbb{R}^d)}\mathcal{F} \leq 0$。特别地， $\pi_\star \in \mathrm{BW}(\mathbb{R}^d)$（$\pi_\star$ 不能属于 $\mathcal{D}$ 否则$\mathcal{F}(\pi_\star) = +\infty$）。因此，我们还有 $\mathcal{F}(\pi_\star) - \min_{\mathrm{BW}(\mathbb{R}^d)}\mathcal{F} \geq 0$，这意味着 $\mathcal{F}(\pi_\star) - \min_{\mathrm{BW}(\mathbb{R}^d)}\mathcal{F} = 0$。最后， $\pi_\star$ 是问题 (1) 的极小化器。

为完成证明，注意到 $(p_k)_{k\geq 0}$ 弱收敛于 $\pi_\star$，且这些分布均为高斯分布。因此，该收敛实际上发生在 Wasserstein 距离下。 □

E.3 定理证明 5.4

证明。 由于 $\mathcal{F}(\hat{\pi}) \leq \mathcal{F}(p_{k+1})$，由于 $\hat{\pi}$ 在高斯分布中实现了 $\mathcal{F}$ 的最小值，我们可以迭代不等式 38，从而得到

$$W_2^2(p_N, \hat{\pi}) \leq \exp(-N\alpha\eta)\, W_2^2(p_0, \hat{\pi}).$$

因此，在选择了

$$\eta = \frac{1}{\beta}, \quad \text{and} \quad N \gtrsim \frac{1}{\alpha\eta}\log\frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2} \asymp \frac{\beta}{\alpha}\log\frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2},$$

我们得到保证 $\alpha W_2^2(p_N, \hat{\pi}) \le \varepsilon^2$.

现在，为了得到 KL 散度意义下的保证，我们用分布 p_N 重新初始化算法，并应用定理 5.2的凸性结论。采用相同的 N 和 η，并假设 ε 足够小，我们可以应用定理5.2，从而得到如下保证

$$\mathcal{F}(p_{2N}) - \mathcal{F}(\hat{\pi}) \le \frac{W_2^2(p_N, \hat{\pi})}{2\eta N} \le \frac{\varepsilon^2}{2\alpha\eta N} \lesssim \frac{\varepsilon^2}{\log \frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2}} \lesssim \varepsilon^2 ,$$

由此证明我们所需的结果。 □

E.4 定理 5.5 的证明

首先，我们需要一个引理。

引理 E.1。 设 $\mu_0, \mu_1 \in \mathsf{BW}(\mathbb{R}^d)$ 满足 $\Sigma_{\mu_0}, \Sigma_{\mu_1} \succeq \beta^{-1} I$。那么，若 (X_0, X_1) 与 (μ_0, μ_1) 对于 W_2 距离而言是最优耦合，则有

$$\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 \le 20\beta^2\, W_2^2(\mu_0, \mu_1) .$$

证明过程如下。首先，我们应用三角不等式和柯西–施瓦茨不等式，将左端分解为两项，并分别进行控制。对于第一项，我们利用 $\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)$ 的 Lipschitz 连续性，这样做是可行的，因为 $\Sigma_{\mu_1}^{-1} \preceq \beta I$。对于第二项，我们将利用引理 B.1 和引理 B.3，推导出一个关于 $\mathsf{KL}(\mu_0 \,\|\, \mu_1)$ 的界，然后再将其进一步界定为关于 $W_2^2(\mu_0, \mu_1)$ 的界。结合这些界，即可得到我们所需的结果。

证明。 应用三角不等式和柯西–施瓦茨不等式，我们得到

$$\begin{aligned}\frac{1}{2}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 \le{}& \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 \\ &+ \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 .\end{aligned}$$

对于第一项，我们注意到，由于 $\Sigma_{\mu_1}^{-1} \preceq \beta I$ 根据假设，我们有

$$\begin{aligned}\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 &= \mathbb{E}\|\Sigma_{\mu_1}^{-1}(X_1 - X_0)\|^2 && \text{(by Equation (28))} \\ &\le \beta^2\,\mathbb{E}\|X_1 - X_0\|^2 && \text{(since } \Sigma_{\mu_1}^{-1} \preceq \beta I\text{)} \\ &= \beta^2\, W_2^2(\mu_0, \mu_1) , && (43)\end{aligned}$$

其中，最后一个等式成立，是因为根据假设，$(X_0, X_1) \sim (\mu_0, \mu_1)$ 是最优耦合的。现在，我们对第二项进行界定。定义泛函 $\mathcal{V}_1, \mathcal{F}_1 : \mathsf{BW}(\mathbb{R}^d) \to \mathbb{R}$ 使得

$$\begin{aligned}\mathcal{V}_1(\mu) &:= -\int \log \mu_1(x)\, \mathrm{d}\mu(x) , \\ \mathcal{F}_1(\mu) &:= \mathcal{V}_1(\mu) + \mathcal{H}(\mu) .\end{aligned}$$

注意，根据公式（28），$\nabla_{\mathsf{BW}}\mathcal{V}_1(\mu) = -\nabla \log \mu_1 = -\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)$，从而

$$\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu) = \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu) + \nabla_{\mathsf{BW}}\mathcal{H}(\mu) = \nabla_{\mathsf{BW}}\mathcal{H}(\mu) - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1) .$$

此外，我们还注意到

$$\mathsf{KL}(\mu \,\|\, \mu_1) = \mathcal{F}_1(\mu) - \mathcal{F}_1(\mu_1) . \tag{44}$$

因此，上面我们希望控制的第二项可以解释为$\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_0)$的平方范数。我们将证明 $\mathcal{F}_1$ 是光滑的，这将使我们能够根据优化中的下降引理，将梯度平方范数界定为 $\mathcal{F}_1(\mu_0) - \mathcal{F}_1(\mu_1) = \mathsf{KL}$ 散度$(\mu_0 \,\|\, \mu_1)$ 的倍数。

令 $\gamma := c^{-}\beta,^{1}$ w 此处选择 $c \in (0,1)$，使其满足 $c \le (1-c)^2$。定义随机变量 X_0' 如下。 s:

$$X_0' := X_0 - \frac{1}{\gamma}\,\nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_0)[X_0]$$

$$
\begin{aligned}
&= X_0 - \frac{1}{\gamma}\left(\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0) - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)\right)[X_0] \\
&= X_0 - \frac{1}{\gamma}\left(-\Sigma_{\mu_0}^{-1}(X_0 - m_{\mu_0}) + \Sigma_{\mu_1}^{-1}(X_0 - m_{\mu_1})\right) && \text{(by Equation (28))} \\
&= \underbrace{\left(I + \frac{1}{\gamma}\Sigma_{\mu_0}^{-1} - \frac{1}{\gamma}\Sigma_{\mu_1}^{-1}\right)}_{:=M_0} X_0 + \frac{1}{\gamma}\left(-\Sigma_{\mu_0}^{-1}m_{\mu_0} + \Sigma_{\mu_1}^{-1}m_{\mu_1}\right).
\end{aligned}
$$

Let $\mu_0' \coloneqq$ 分布(X_0')。由于我们有 $0 \preceq \Sigma_{\mu_0}^{-1}, \Sigma_{\mu_1}^{-1} \preceq \beta I = c\gamma I$ 根据假设，我们有

$$
M_0 = I + \frac{1}{\gamma}\Sigma_{\mu_0}^{-1} - \frac{1}{\gamma}\Sigma_{\mu_1}^{-1} \succeq I - \frac{1}{\gamma}\Sigma_{\mu_1}^{-1} \succeq (1-c)\, I \succeq 0\,,
$$

因此 X_0' 等于的凸函数的梯度 X_0。因此，根据布雷尼尔定理，我们得出（X_0， X_0'） $\sim$ （μ_0， μ_0'）关于距离是最优耦合的。因此，根据应用于势函数$\mathcal{V}_1$的引理

$$
\begin{aligned}
\mathcal{V}_1(\mu_0') - \mathcal{V}_1(\mu_0) &\le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu_0)[X_0], X_0' - X_0\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0' - X_0\|^2 && \text{(since } -\nabla^2 \log \mu_0 \preceq \beta I\text{)} \\
&= -\mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0], X_0' - X_0\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0' - X_0\|^2\,. && (45)
\end{aligned}
$$

此外，我们注意到，由于 $\beta = c\gamma \le (1-c)^2\,\gamma$，我们有】【。

$$
\Sigma_{\mu_0'} = M_0 \Sigma_{\mu_0} M_0 \succeq (1-c)^2\,\Sigma_{\mu_0} \succeq \frac{(1-c)^2}{\beta}\, I \succeq \frac{1}{\gamma}\, I\,.
$$

这意味着 $\Sigma_{\mu_0'}^{-1}$， $\Sigma_{\mu_0}^{-1} \preceq \gamma I$。因此，我们也可以应用引理 B.3的测地线光滑性不等式来得到

$$
\mathcal{H}(\mu_0') - \mathcal{H}(\mu_0) \le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0], X_0' - X_0\rangle + \frac{\gamma}{2}\,\mathbb{E}\|X_0' - X_0\|^2\,. \tag{46}
$$

因此，结合等式（44）与不等式45以及不等式46，我们得到 at

$$
\begin{aligned}
-\text{KL 散度}(\mu_0 \,\|\, \mu_1) &\le \text{KL 散度}(\mu_0' \,\|\, \mu_1) - \text{KL 散度}(\mu_0 \,\|\, \mu_1) && (\text{since } \mathsf{KL}(\mu_0' \,\|\, \mu_1) \ge 0) \\
&= \mathcal{F}_1(\mu_0') - \mathcal{F}_1(\mu_0) && \text{(by Equation (44))} \\
&= [\mathcal{V}_1(\mu_0') - \mathcal{V}_1(\mu_0)] + [\mathcal{H}(\mu_0') - \mathcal{H}(\mu_0)] \\
&\le -\mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0], X_0' - X_0\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0' - X_0\|^2 && \text{(by Inequality 45)} \\
&\quad + \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0], X_0' - X_0\rangle + \frac{\gamma}{2}\,\mathbb{E}\|X_0' - X_0\|^2 && \text{(by Inequality 46)} \\
&= \left(-\frac{1}{\gamma} + \frac{\beta}{2\gamma^2} + \frac{1}{2\gamma}\right)\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 && \text{(definition of } X_0') \\
&= -\frac{1-c}{2\gamma}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2\,. && (47)
\end{aligned}
$$

为了对该不等式的左侧进行界定，我们再次将引理B.1应用于势函数 $\mathcal{V}_1$，并将引理 B.3 应用于$\mathcal{H}$，从而得到

$$
\begin{aligned}
\mathsf{KL}(\mu_0 \,\|\, \mu_1) &= \mathcal{F}_1(\mu_0) - \mathcal{F}_1(\mu_1) \\
&= [\mathcal{V}_1(\mu_0) - \mathcal{V}_1(\mu_1)] + [\mathcal{H}(\mu_0) - \mathcal{H}(\mu_1)] \\
&\le \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{V}(\mu_1)[X_1], X_0 - X_1\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0 - X_1\|^2 && \text{(by Lemma B.1 since } -\nabla^2 \log \mu_1 \preceq \beta I\text{)} \\
&\quad + \mathbb{E}\langle \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1], X_0 - X_1\rangle + \frac{\beta}{2}\,\mathbb{E}\|X_0 - X_1\|^2 && \text{(by Lemma B.3 since } \Sigma_{\mu_0}^{-1}, \Sigma_{\mu_1}^{-1} \preceq \beta I\text{)}
\end{aligned}
$$

$$= \beta\,\mathbb{E}\|X_0 - X_1\|^2 \qquad (\text{since } \nabla_{\mathsf{BW}}\mathcal{V}_1(\mu_1) + \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1) = \nabla_{\mathsf{BW}}\mathcal{F}_1(\mu_1) = 0)$$
$$= \beta\, W_2^2(\mu_0, \mu_1)\,. \tag{48}$$

最后，选择 $c = \frac{1}{3}$ 使得 $c \le (1-c)^2$，并结合上述不等式，我们得到

$$\begin{aligned}\frac{1}{2}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0]\|^2 &\le \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_1] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 \\ &\quad + \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{H}(\mu_0)[X_0] - \nabla_{\mathsf{BW}}\mathcal{H}(\mu_1)[X_0]\|^2 \\ &\le \beta^2\, W_2^2(\mu_0,\mu_1) + \frac{2\gamma}{1-c}\,\mathsf{KL}(\mu_0 \,\|\, \mu_1) \\ &\qquad (\text{by Inequality 43 and Inequality 47}) \\ &\le 10\beta^2\, W_2^2(\mu_0,\mu_1)\,. \qquad (\text{by Inequality 48})\end{aligned}$$

重新整理后，即得所需结果。 □

有了这一结果，我们就可以证明所需的驻点保证。

证明。 设$(X_k, X_{k+\frac{1}{2}}) \sim (p_k, p_{k+\frac{1}{2}})$和$(X_{k+\frac{1}{2}}, X_{k+1}) \sim (p_{k+\frac{1}{2}}, p_k)$在 W_2距离下是最优耦合的。与引理5.1 的证明一样，注意到根据构造，

$$\frac{X_k - X_{k+1}}{\eta} = \nabla_{\mathsf{BW}}\mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\,.$$

应用引理C.2，可得

$$\mathbb{E}\|X_{k+1} - X_k\|^2 \le -2\eta\,\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)]\,.$$

将此不等式逐项累加，可得

$$\begin{aligned}\min_{k\in\{0,\dots,N-1\}} \mathbb{E}\|X_{k+1} - X_k\|^2 &\le \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\|X_{k+1} - X_k\|^2 \\ &\le -\frac{2\eta}{N}\sum_{k=0}^{N-1}\mathbb{E}[\mathcal{F}(p_{k+1}) - \mathcal{F}(p_k)] \\ &= -\frac{2\eta}{N}\,\mathbb{E}[\mathcal{F}(p_N) - \mathcal{F}(p_0)] \\ &\le \frac{2\eta\Delta}{N}\,. \end{aligned} \tag{49}$$

现在，设$(X_k, X_{k+1}^\star) \sim (p_k, p_{k+1})$在 W_2 距离下是最优耦合的。根据推论D.2，有$\Sigma_k^{-1} \preceq \beta I$ 对所有 k均成立，这意味着我们可以应用引理E.1 ，取 $\mu_0 = p_k$ 和 $\mu_1 = p_{k+1}$ ，从而得到

$$\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{H}(p_k)[X_k] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}^\star]\right\|^2 \le 20\beta^2\, W_2^2(p_k, p_{k+1})\,. \tag{50}$$

此外，我们还有

$$\begin{aligned}\mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}^\star] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\right\|^2 &= \mathbb{E}\left\|\Sigma_{k+1}^{-1}\,(X_{k+1}^\star - X_{k+1})\right\|^2 \\ &\le \beta^2\,\mathbb{E}\left\|X_{k+1}^\star - X_{k+1}\right\|^2 \\ &\le 2\beta^2\,\mathbb{E}\left\|X_{k+1}^\star - X_k\right\|^2 + 2\beta^2\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\ &= 2\beta^2\, W_2^2(p_k, p_{k+1}) + 2\beta^2\,\mathbb{E}\|X_{k+1} - X_k\|^2\,. \end{aligned} \tag{51}$$

考虑到这些不等式，我们得到：

$$\frac{1}{3}\,\|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|_{p_k}^2 = \frac{1}{3}\,\mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}}\mathcal{H}(p_k)[X_k]\|^2$$

$$
\begin{aligned}
&\le \mathbb{E}\|\nabla_{\mathsf{BW}}\mathcal{V}(p_k)[X_k] + \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\|^2 + \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{H}(p_k)[X_k] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X^\star_{k+1}]\right\|^2 \\
&\quad + \mathbb{E}\left\|\nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X^\star_{k+1}] - \nabla_{\mathsf{BW}}\mathcal{H}(p_{k+1})[X_{k+1}]\right\|^2 && \text{(by triangle inequality)} \\
&\le \frac{1}{\eta^2}\,\mathbb{E}\|X_{k+1} - X_k\|^2 + 22\,\beta^2 W_2^2(p_k, p_{k+1}) + 2\beta^2\,\mathbb{E}\|X_{k+1} - X_k\|^2 \\
& && \text{(by Inequality 50 and Inequality 51)} \\
&\le \left(\frac{1}{\eta^2} + 24\beta^2\right)\mathbb{E}\|X_{k+1} - X_k\|^2 && \text{(since } (X_k, X_{k+1}) \text{ is a coupling of } (p_k, p_{k+1})\text{)} \\
&\le \frac{25}{\eta^2}\,\mathbb{E}\|X_{k+1} - X_k\|^2\,. && \text{(since } \beta \le \eta^{-1}\text{)}
\end{aligned}
$$

结合上述结果与不等式 49，可得

$$
\min_{k\in\{0,\dots,N-1\}} \|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|_{p_k}^2 \le \min_{k\in\{0,\dots,N-1\}} \frac{75}{\eta^2}\,\mathbb{E}\|X_{k+1} - X_k\|^2 \le \frac{150\Delta}{\eta N}\,.
$$

最后，取 $\eta = \frac{1}{\beta}$ 和 $N \ge \frac{150\beta\Delta}{\varepsilon^2}$，我们得到】【：】【“】【

$$
\min_{k\in\{0,\dots,N-1\}} \|\nabla_{\mathsf{BW}}\mathcal{F}(p_k)\|_{p_k}^2 \le \varepsilon^2\,,
$$

如所需。 □

F 噪声算法收敛速率的证明

我们再次利用引理 5.1 来获得所需的收敛速率。首先，我们必须证明引理 σ_k 中给出的随机 FB–GVI 的界 5.6。

F.1 引理的证明 5.6

证明。 令 $\mu = \mathcal{N}(m\ \Sigma)$ 是 BW($\mathbb{R}^d$) 的一个元素。我们首先注意到，如果 $X \sim \mu$，那么通过分部积分 s,

$$
\begin{aligned}
\Sigma\,\mathbb{E}\nabla^2 V(X) &= \Sigma \int \nabla^2 V\,\mathrm{d}\mu \\
&= -\Sigma \int \nabla\mu \otimes \nabla V && \text{(integration by parts)} \\
&= -\Sigma \int \nabla \ln\mu \otimes \nabla V\,\mathrm{d}\mu \\
&= \int (x - m) \otimes \nabla V\,\mathrm{d}\mu(x) && \text{(since } -\Sigma\,\nabla \ln\mu(x) = x - m\text{)} \\
&= \mathbb{E}[(X - m) \otimes \nabla V(X)]\,. && (52)
\end{aligned}
$$

因此，

$$
\begin{aligned}
\langle \mathbb{E}\nabla^2 V(X), \Sigma\rangle &= \langle \mathbb{E}[\Sigma^{-1}\,(X - m) \otimes \nabla V(X)], \Sigma\rangle && \text{(by Equation (52))} \\
&= \mathbb{E}\langle \Sigma^{-1}\,(X - m) \otimes \nabla V(X), \Sigma\rangle && \text{(linearity of expectation and trace)} \\
&= \mathbb{E}\langle \nabla V(X), X - m\rangle\,. && \text{(cyclicity of trace)}
\end{aligned}
$$

现在，令 $(X_k, Z) \sim (p_k, \hat{}\pi)$ 在 W_2 距离下达到最优耦合，且与 $\hat{}\ X_k$ 相互独立。另回顾 Brascamp–Lieb 不等式（Brascamp 和 Lieb，1976）：若 μ 是 $\mathbb{R}^d$ 上的一个测度，其密度为 $\mu \propto \exp(-W)$，其中 W 是二次连续可微且严格凸的，则对任意光滑测试函数 f，都有： $\mathbb{R}^d \to \mathbb{R}$it holds that Var$_\mu(f) \le \mathbb{E}\langle \nabla f, (\nabla^2 W)^{-1}\,\nabla f\rangle$。特别地，若我们取 $f = \langle \nabla V, e\rangle$ 为单位向量 e 且$\mu = p_k$，则可得 Var$_{p_k}\langle \nabla V, e\rangle \le \mathbb{E}_{p_k}\langle e, \nabla^2 V\,\Sigma_k\,\nabla^2 V\,e\rangle$。将此不等式对 e 取遍 $\mathbb{R}^d$的一组标准正交基后求和，可得

$$
\mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 \le \mathbb{E}_{p_k}\langle [\nabla^2 V]^2, \Sigma_k\rangle\,.
$$

因此，我们得到

$$
\begin{aligned}
\frac{1}{2}\,\sigma_k^2 &\le \mathbb{E}\|(\nabla^2 V(\hat{X}_k) - \mathbb{E}_{p_k}\nabla^2 V)(X_k - m_k)\|^2 + \mathbb{E}\|\nabla V(\hat{X}_k) - \mathbb{E}_{p_k}\nabla V\|^2 && \text{(by triangle inequality)}\\
&= \left\langle \mathbb{E}_{p_k}[(\nabla^2 V - \mathbb{E}_{p_k}\nabla^2 V)^2], \Sigma_k \right\rangle + \mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 && (\text{since } X_k \perp\!\!\!\perp \hat{X}_k)\\
&= \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle - \langle \mathbb{E}_{p_k}[\nabla^2 V]^2, \Sigma_k\rangle + \mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 \\
&\le \mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle + \mathbb{E}_{p_k}\|\nabla V - \mathbb{E}_{p_k}\nabla V\|^2 && (\text{since } \langle \mathbb{E}_{p_k}[(\nabla^2 V)^2], \Sigma_k\rangle \ge 0)\\
&\le 2\,\mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\, \nabla^2 V\rangle && \text{(by Brascamp–Lieb)}\\
&\le 2\beta\,\mathbb{E}_{p_k}\langle \nabla^2 V, \Sigma_k\rangle && (\text{since } \nabla^2 V \preceq \beta I \text{ and } \nabla^2 V, \Sigma_k \succeq 0)\\
&= 2\beta\,\mathbb{E}\langle \nabla V(X_k), X_k - m_k\rangle && (\text{by Equation (52)})\\
&= 2\beta\,\underbrace{\mathbb{E}\langle \nabla V(Z), Z - \hat{m}\rangle}_{\mathsf{err}_1} + 2\beta\,\underbrace{\mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle}_{\mathsf{err}_2}\\
&\quad + 2\beta\,\underbrace{\mathbb{E}\langle \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle}_{\mathsf{err}_3} + 2\beta\,\underbrace{\mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), Z - \hat{m}\rangle}_{\mathsf{err}_4}.
\end{aligned}
$$

现在，我们得到如下结论：

$$
\begin{aligned}
\mathsf{err}_1 &= \mathbb{E}\langle \nabla V(Z), Z - \hat{m}\rangle = \langle \mathbb{E}\nabla^2 V(Z), \hat{\Sigma}\rangle = \mathrm{Tr}(I)\\
&\qquad\qquad \text{(by Equation (52) and the stationarity conditions in (29))}\\
&= d\,,\\
\mathsf{err}_2 &= \mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle\\
&\le \frac{1}{2\beta}\,\mathbb{E}\|\nabla V(X_k) - \nabla V(Z)\|^2 + \frac{\beta}{2}\,\mathbb{E}\|(X_k - m_k) - (Z - \hat{m})\|^2 && \text{(Young's inequality)}\\
&\le \beta\,\mathbb{E}\|X_k - Z\|^2 && (\text{since } \nabla V \text{ is } \beta\text{-Lipschitz})\\
&= \beta\, W_2^2(\mu_k, \hat{\pi})\,, && (\text{since } (X_k, Z) \text{ are optimally coupled})\\
\mathsf{err}_3 &= \mathbb{E}\langle \nabla V(Z), (X_k - m_k) - (Z - \hat{m})\rangle\\
&\le \frac{1}{4\beta}\,\mathbb{E}\|\nabla V(Z)\|^2 + \beta\,\mathbb{E}\|(X_k - m_k) - (Z - \hat{m})\|^2 && \text{(Young's inequality)}\\
&\le \frac{1}{4\beta}\,\mathbb{E}\langle \nabla^2 V(Z)^2, \hat{\Sigma}\rangle + \beta\, W_2^2(\mu_k, \hat{\pi}) && \text{(Brascamp–Lieb, optimal coupling of } (X_k, Z))\\
&\le \frac{d}{4} + \beta\, W_2^2(\mu_k, \hat{\pi})\,, && (\text{since } \mathbb{E}_{\hat{\pi}}\nabla^2 V = \hat{\Sigma}^{-1} \text{ by Equation (29) and } \nabla^2 V \preceq \beta I)\\
\mathsf{err}_4 &= \mathbb{E}\langle \nabla V(X_k) - \nabla V(Z), Z - \hat{m}\rangle\\
&\le \frac{\mathrm{Tr}(\hat{\Sigma})}{d}\,\mathbb{E}\|\nabla V(X_k) - \nabla V(Z)\|^2 + \frac{d}{4\,\mathrm{Tr}(\hat{\Sigma})}\,\mathbb{E}\|Z - \hat{m}\|^2 && \text{(Young's inequality)}\\
&\le \frac{\beta^2\,\mathrm{Tr}(\hat{\Sigma})}{d}\,\mathbb{E}\|X_k - Z\|^2 + \frac{d}{4\,\mathrm{Tr}(\hat{\Sigma})}\,\mathrm{Tr}(\hat{\Sigma}) && (\text{since } \nabla V \text{ is } \beta\text{-Lipschitz})\\
&\le \frac{\beta^2\,\mathrm{Tr}(\hat{\Sigma})}{d}\,W_2^2(\mu_k, \hat{\pi}) + \frac{d}{4}\,.
\end{aligned}
$$

结合这些结果，我们得到：

$$
\sigma_k^2 \le 4\beta\sum_{i=1}^{4}\mathsf{err}_i \le 6\beta d + \left(8\beta^2 + \frac{4\beta^3\,\mathrm{Tr}(\hat{\Sigma})}{d}\right) W_2^2(\mu_k, \hat{\pi}) \le 6\beta d + 12\beta^3\lambda_{\max}(\hat{\Sigma})\, W_2^2(\mu_k, \hat{\pi})\,.
$$

(since $\hat{\Sigma}^{-1} = \mathbb{E}_{\hat{\pi}}\nabla^2 V \preceq \beta I$ so $\lambda_{\max}(\hat{\Sigma}) \ge 1/\beta$)

注意，在强凸情形下，根据公式（29），我们得到

$$
\lambda_{\max}(\hat{\Sigma}) = \lambda_{\max}(\mathbb{E}_{\hat{\pi}}[\nabla^2 V]^{-1}) \le \frac{1}{\alpha}\,,
$$

因此，该界简化为

$$\sigma_k^2 \le 6\beta d + \frac{12\beta^3}{\alpha} W_2^2(\mu_k, \hat{\pi})\,.$$

至此，我们的证明完成。 □

F.2 利用对 σ_k 的界的一步不等式

我们将引理 5.6 的误差界与引理 5.1 的一步不等式结合起来，其中 $\nu = \hat{\pi}$ 和 $\eta \le \frac{1}{2\beta}$。这给出了不等式

$$\begin{aligned}
\mathbb{E}W_2^2(p_{k+1}, \hat{\pi}) &\le (1-\alpha\eta)\mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 2\eta^2\,\mathbb{E}\sigma_k^2 \\
&\le \left(1 - \alpha\eta + 24\beta^3\eta^2\lambda_{\max}(\hat{\Sigma})\right)\mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d \\
&\le \exp\left(-\alpha\eta + 24\beta^3\eta^2\lambda_{\max}(\hat{\Sigma})\right)\mathbb{E}W_2^2(p_k, \hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d\,. \qquad (53)
\end{aligned}$$

F.3 定理5.7 的证明

证明。 定义 $c := 24\beta^3\lambda_{\max}(\hat{\Sigma})$。由于 V 根据假设是凸的，我们可以在不等式 53 中取 $\alpha = 0$ 以得到

$$2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) \le e^{c\eta^2}\,\mathbb{E}W_2^2(p_k, \hat{\pi}) - \mathbb{E}W_2^2(p_{k+1}, \hat{\pi}) + 12\beta\eta^2 d\,.$$

定义 $S_N(\eta) := \sum_{k=1}^N e^{-kc\eta^2}$。然后我们发现

$$\begin{aligned}
\sum_{k=0}^{N-1} 2\eta\, e^{-(k+1)c\eta^2}\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) &\le \sum_{k=0}^{N-1} e^{-(k+1)c\eta^2}\left(e^{c\eta^2}\,\mathbb{E}W_2^2(p_k, \hat{\pi}) - \mathbb{E}W_2^2(p_{k+1}, \hat{\pi}) + 12\beta\eta^2 d\right) \\
&= W_2^2(p_0, \hat{\pi}) - e^{-Nc\eta^2}\,\mathbb{E}W_2^2(p_N, \hat{\pi}) + 12\beta\eta^2 d\sum_{k=0}^{N-1} e^{-(k+1)c\eta^2} \\
&\le W_2^2(p_0, \hat{\pi}) + 12\beta\eta^2 d S_N(\eta)\,.
\end{aligned}$$

设 $\bar{p}$ 从 $\{p_k\}_{k=1}^N$ 中随机抽取，选择 p_k 的概率与 $e^{-kc\eta^2}$ 成正比。于是我们有。

$$\begin{aligned}
\mathbb{E}\mathcal{F}(\bar{p}) - \mathcal{F}(\hat{\pi}) &= \frac{1}{2\eta S_N(\eta)}\sum_{k=0}^{N-1} 2\eta\, e^{-(k+1)c\eta^2}\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) \\
&\le \frac{1}{2\eta S_N(\eta)}\left(W_2^2(p_0, \hat{\pi}) + 12\beta\eta^2 d S_N(\eta)\right) \\
&= \frac{W_2^2(p_0, \hat{\pi})}{2\eta S_N(\eta)} + 6\beta\eta d\,.
\end{aligned}$$

现在，我们注意到

$$S_N(\eta) = \sum_{k=1}^N e^{-kc\eta^2} \ge \sum_{k=1}^{N\wedge(c\eta^2)^{-1}} e^{-kc\eta^2} \ge \sum_{k=1}^{N\wedge(c\eta^2)^{-1}} e^{-1} \ge \frac{N \wedge \lfloor (c\eta^2)^{-1}\rfloor}{e}\,.$$

因此，我们得到不等式

$$\begin{aligned}
\mathbb{E}\left[\min_{k\in\{1,\dots,N\}}\mathcal{F}(p_k)\right] - \mathcal{F}(\hat{\pi}) &\le \mathbb{E}\mathcal{F}(\bar{p}) - \mathcal{F}(\hat{\pi}) \\
&\le \frac{W_2^2(p_0, \hat{\pi})}{2\eta S_N(\eta)} + 6\beta\eta d
\end{aligned}$$

$$\leq \frac{2W_2^2(p_0,\hat{\pi})}{\eta\left(N \wedge \lfloor (c\eta^2)^{-1} \rfloor\right)} + 6\beta\eta d$$
$$\lesssim \frac{W_2^2(p_0,\hat{\pi})}{\eta N} + c\eta W_2^2(p_0,\hat{\pi}) + \beta\eta d\,.$$

因此，取

$$\eta \asymp \frac{\varepsilon^2}{cW_2^2(p_0,\hat{\pi}) \vee \beta d} \asymp \frac{\varepsilon^2}{\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(p_0,\hat{\pi}) \vee \beta d}\,,$$
$$N \gtrsim \frac{W_2^2(p_0,\hat{\pi})}{\eta\varepsilon^2} \asymp \frac{W_2^2(p_0,\hat{\pi})}{\varepsilon^4}\left(\beta^3 \lambda_{\max}(\hat{\Sigma})\, W_2^2(p_0,\hat{\pi}) \vee \beta d\right),$$

我们得到保证 $\mathbb{E}\big[\min_{k\in\{1,\ldots,N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \leq \varepsilon^2$。 □

F.4 定理5.8 的证明

证明。 在强凸情形下，当 $0 \prec \alpha I \preceq \nabla^2 V$时，我们有特征值保证 $\lambda_{\max}(\hat{\Sigma}) \leq \frac{1}{\alpha} I$，因为 $\mathbb{E}_\pi \nabla^2 V = \hat{\Sigma}^{-1}$。根据（29）。因此，在假设 $\eta \leq \frac{\alpha^2}{48\beta^3}$成立的情况下，不等式 53 意味着——

$$\mathbb{E}W_2^2(p_{k+1},\hat{\pi}) \leq \exp\Big(-\alpha\eta + \frac{24\beta^3\eta^2}{\alpha}\Big)\,\mathbb{E}W_2^2(p_k,\hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d$$
$$\leq \exp\Big(-\frac{\alpha\eta}{2}\Big)\,\mathbb{E}W_2^2(p_k,\hat{\pi}) - 2\eta\left(\mathbb{E}\mathcal{F}(p_{k+1}) - \mathcal{F}(\hat{\pi})\right) + 12\beta\eta^2 d\,.$$

由于 $\mathcal{F}(\hat{\pi}) \leq \mathcal{F}(p_{k+1})$，我们可以迭代这一不等式，从而得到

$$\mathbb{E}W_2^2(p_N,\hat{\pi}) \leq \exp\Big(-\frac{N\alpha\eta}{2}\Big)\, W_2^2(p_0,\hat{\pi}) + \frac{24\beta\eta d}{\alpha}\,.$$

因此，取

$$\eta \asymp \frac{\varepsilon^2}{\beta d}\,, \quad \text{and} \quad N \gtrsim \frac{1}{\alpha\eta}\log\frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2} \asymp \frac{\beta d}{\alpha\varepsilon^2}\log\frac{\alpha W_2^2(p_0,\hat{\pi})}{\varepsilon^2}\,,$$

我们得到保证 $\alpha\,\mathbb{E}W_2^2(p_N, \hat{\pi}) \leq \varepsilon^2$。现在，对于 KL 散度中的保证，我们使用分布 p_N 重新初始化算法，并应用定理 5.7的凸结果。该论证受到Durmus 等人 (2019) 的启发。假设 ε 足够小，我们得到

$$c\,\mathbb{E}W_2^2(p_N,\hat{\pi}) \leq \frac{c\varepsilon^2}{\alpha} \leq \beta d\,,$$

这意味着，对于上述 η 的选择，我们有

$$\eta \asymp \frac{\varepsilon^2}{\beta d} \asymp \frac{\varepsilon^2}{c\,\mathbb{E}W_2^2(p_N,\hat{\pi}) \vee \beta d}\,.$$

此外，对于我们选择的 N，我们有

$$\frac{\mathbb{E}W_2^2(p_N,\hat{\pi})}{\varepsilon^4}\left(cW_2^2(p_0,\hat{\pi}) \vee \beta d\right) \leq \frac{\beta d}{\alpha\varepsilon^2} \lesssim N\,.$$

因此，应用定理 5.7 ，并采用我们选择的步长 η 和迭代次数 N，可得 at

$$\mathbb{E}\big[\min_{k\in\{1,\ldots,2N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi}) \leq \mathbb{E}\big[\min_{k\in\{N+1,\ldots,2N\}} \mathcal{F}(p_k)\big] - \mathcal{F}(\hat{\pi})$$
$$\lesssim \mathbb{E}\Big[\frac{W_2^2(p_N,\hat{\pi})}{\eta N} + c\eta\, W_2^2(p_N,\hat{\pi}) + \beta\eta d\Big]$$
$$\lesssim \varepsilon^2\,,$$

从而证明了我们所需的结果。 □