Optimal Transport in Machine Learning

3rd ed, Nov. 30th 2017 Y. Chen

CONTENTS

Some Notation

OT: Optimal transport
PM: Probability measures
W: Wasserstein
PGA: Principle Geodesic Analysis

Statistic

NIPS（2005-2017）：31篇

ICML（2014-2017）：21篇

JMLR（2004-2016）：5篇

PMLR only（2017）：3篇

AISTATS et.al（2014-2017）：5篇

arXiv（2015-2017）：6篇

Development（Time Order）

2004-2010

分类理论
Lipchitz 函数与 1-W

用作度量
p-W 与噪声，ROC 曲率

（收敛性）p-W 与 SGD

2012-2014

用作损失函数
2-W 以及离散 p-W 与聚类，Sinkhorn距离与分类问题，2-W 与图半监督，离散 2-W 与领域适应

用作度量
PV，Ricci 与 Markov Chain，样本 iid

2015

用作损失函数
离散 2-W 与子空间聚类 / 多标签学习

用作度量

（几何）离散 2-W 与度量学习 / barycenter，2-W 与 PGA / PDA，p-W 与数据拓扑（ TDA / lanscape / PH）

（其他）1-W 与 LR 模糊集，Kantorovich 度量与 Markov，IPM 与样本质量，TV 与 PLMC

2016

用作损失函数

（生成式模型）W 与 RBM，2-W 与 RNN

（无监督）离散1-W 与字典学习，离散 OT与music transcription

（有监督）离散 W 与 WMD&SWMD，离散2-W与WDA

用作度量

（几何）离散 OT 与barycenter

（其他）离散 OT 与非线性相关关系

求解 OT

大规模OT，最优映射函数

2017

用作损失函数

（有监督）2-W与单标签分类

（无监督）1-W与噪声预测，离散2-W与多层次聚类 / 联合聚类，MMD与对抗特征匹配

（GAN）JS与GAN，1-W与WGAN，Cramer与CGAN，F距离与MIX+GAN，IPM/MMD与McGAN&Fisher GAN&MMD GAN

（其他）PMD与深度学习

用作度量

（度量收敛性 / 误差）1-W 与扩散Markov / sampler，2-W与NN / LMC / SGLD，IPM与sample质量，IPM与个体预测

（几何）Sliced W kernel与PD， 2-W与不确定曲线（高斯过程）

（其他）W距离与强化学习，W 距离与multiple populations 空间

求解 OT

离散W loss优化，张量平衡与解 W 距离，变分推断与OT，GAN的训练，Sinkhorn距离

Wasserstein Distance

“ W 距离度量弱的收敛性。虽然也有其他度量同样如此，但 W 距离相比其他有特别的优势 (Villani[35, p. 98])：我们不难通过使用连续的边界限制从 W 距离的收敛性中得到更强度量下的收敛性。” W 距离在 ML 中应用最多的是 2-W 距离以及其离散形式。2-W 有自然的几何意义，以及 2-W 距离下定义的 Barycenter 在聚类、数据拓扑分析等等方面有很多应用。

Supervised Learning

1-W

1. Distance-Based Classification with Lipschitz Functions (JMLR 2004)
与 1-W 距离的 K-L 对偶有关，讨论方法是将 Lipschitz 函数空间嵌入到一个度量空间的对偶空间中。

1. Distributionally Robust Logistic Regression (NIPS 2015)
使用了 1-W 距离构建模糊集。在分布式鲁棒优化算法中，我们使用训练集构造一个模糊集 $\mathscr{P}$，该集合是一族分布的集合，该集合包含未知真实分布 $\mathbb{P}$ 有很高的置信度。我们使用 1-W 距离，将 $\mathscr{P}$ 构造为概率分布空间中的以训练数据的经验分布为中心的一个球。

2-W

2. Wasserstein Training of Restricted Boltzmann Machines (NIPS 2016)
RBM。在假设我们知道观测数据之间的一个有意义的度量的情况下，可以利用这一度量定义 RBM 生成的分布与训练样本给出的分布之间的 W 距离。推导出了用模型参数表示的这一距离的梯度。使用了熵正则化的 W 距离。

1. Squared Earth Mover’s Distance-based Loss for Training Deep Neural Networks (arXiv 2017)
单标签分类问题。使用 2-W 作为损失函数，从而能够考虑到不同类别之间的联系。在不需要类别间联系的先验知识的条件下同样可以学习到一个距离矩阵。此算法在强或弱的关联下都有比较好的表现。

DISCRETE 2-W

3. Learning with a Wasserstein Loss (NIPS 2015)
多标签学习。使用离散 2-W 距离度量生成标签和真实标签之间的距离。使用准确 2-W 距离的熵正则化近似来简化计算，并在此基础上运用一种高效的学习算法。

4. Regressive Virtual Metric Learning (NIPS 2015)
使用“虚拟点”的方法进行度量学习。这些虚拟点的选择可以视为离散点上的最优传输问题。使用了离散 2-W 距离。

5. Supervised Word Mover's Distance (NIPS 2016)
WMD 距离是将文档间的距离视为两文档的词向量间的离散 W 距离，是无监督的。这里提出了一种监督式 WMD，基于原先的 WMD 距离加入了一个矩阵和一个权重向量。提出了快速计算损失梯度的方法。

2. Wasserstein Discriminant Analysis (arXiv 2016)
一种新的监督学习方法，通过计算一个合适的线性映射，将高维的数据映射到低维的空间中，从而便于对其进行分类。使用了正规化的离散 2-W 距离来得到这一映射，并使用 Sinkhorn matrix scaling algorithm 来计算。使用 2-W 的好处是可以兼顾全局和局部的约束。本文与 PDA（NIPS 2015）那篇有关系，但那篇只限制在一维空间中。

Unsupervised Learning

1-W

1. Unsupervised Learning by Predicting Noise (ICML 2017)
文中使用的损失函数实际上是 1-W 距离的一种近似。文章提出这种算法比 GAN 更简单、训练更快。

2-W

6. Learning Probability Measures with respect to Optimal Transport Metrics (NIPS 2012)
本文研究对于支持集在一个流形上的 PM，误差使用传输度量时，如何学习此 PM 的问题。此问题与最优传输，最优量子化以及无监督学习三个领域有深刻联系。很多广泛使用的无监督学习算法（如k-means）其实都是在 2-W 度量意义下学习数据的 PM。本文使用这三种领域下的方法研究得到了以下结论：1）在离散集上的 PM 和其 estimates 间距离的一致下界。2）从 emprical measure 到 population measure 的收敛率的新的概率区间。3）从 k-means 生成的 measure 到数据 measure 的收敛率的概率区间。

2. Learning Population-Level Diffusions with Generative Recurrent Networks (ICML 2016)
使用了具有 2-W loss 的 RNN，将扩散过程与 RNN 联系在一起，用于研究群体随时间的动态变化。这里的处理方法是将 2-W 下的二次 loss 视为布朗运动在无穷小时间中将预测值传播到真实值的对数概率，从轨迹线的角度入手。

4. Multilevel Clustering via Wasserstein Means (ICML 2017), Supplementary 🌟
多层次聚类。使用 2-W 距离作为优化目标，其优化问题称为多层次 W 中值问题。通过求不同层次的 W barycenter 进行聚类。

DISCRETE W

1. Domain adaptation with regularized optimal transport (ECML PKDD 2014)
领域适应。使用了离散形式的 2-W 距离。迁移学习是研究分类器从有标签的源域到无标签的目标域进行知识迁移的泛化能力。在领域适应学习中认为源域和目标域不同，其目标是尽可能排除域的差异来度量两种域上的概率分布的相似度。本文利用最优传输距离，提出了 OT-reglab 算法。

7. Differentially private subspace clustering (NIPS 2015)
这里使用的是 2-W 距离的离散形式。子空间聚类是一种以“数据点聚集成的簇近似处于低维线性空间中”为目标的聚类任务。本文提出差分隐私子空间聚类的两种算法，基于 2-W 和矩阵范数定义了一种子空间之间的距离，构造损失函数。

8. Optimal spectral transportation with application to music transcription (NIPS 2016)
音乐标注。使用了离散形式的 OT 距离作为优化目标。

2. Fast Dictionary Learning with a Smoothed Wasserstein Loss (AISTATS 2016)
字典学习。使用了离散形式的 1-W 距离，以及熵正则化。将其 legendre 变换后的对偶形式作为损失函数。

3. Co-clustering through Optimal Transport (ICML 2017) 🌟
联合聚类。从变分推断的角度来看 OT 问题的求解，从而将其应用在联合聚类中。使用了熵正规化的离散 2-W 度量。还介绍此算法的核化形式，使用了 Gromov-Wasserstein 度量。

Semi-supervised Learning

5. Wasserstein Propagation for Semi-Supervised Learning (ICML 2014)
使用了 2-W 距离以及定理[Villani (2003), Theorem 2.18]中 2-W 距离的累积分布函数 (CDF) 的形式，用此距离作为优化目标。文中介绍的 W 传播方法比经典的基于图的半监督学习的标签传播方法高效，并可以扩展到流形上。

Reinforcement Learning

6. A Distributional Perspective on Reinforcement Learning (ICML 2017)
研究强化学习中 agent 所接收的随机返回值的分布。提到了对于一个固定的 policy，作用在概率分布上的 Bellman operator 在 W 度量下是一个压缩映射。

GAN

7. Wasserstein GAN (arXiv 2017), Wasserstein Generative Adversarial Networks (ICML 2017)

使用 1-W 距离的 K-L 对偶形式作为生成器和判别器对抗学习的目标函数。判别器被限制在 Lipchitz 函数空间，因此本文使用 weight clipping 方法，将判别器的参数限制在一个确定的闭区间中。但此方法存在问题，区间过大过小都会影响学习效果。

3. Improved Training of Wasserstein GANs (arXiv 2017)
由于 Lipchitz 函数即是梯度的范数处处小于1的函数，这里在目标函数中加入梯度惩罚来限制判别器，并只限制其在样本数据和生成数据之间点上的梯度。使用此方法后 WGAN 的表现进一步提高。

Geometry / Topology

BARYCENTER

8. Fast Computation of Wasserstein Barycenters (ICML 2014)
介绍了离散形式的 p-W 距离，三种算法。第三种算法使用了正则化的主问题和光滑化的对偶问题。可用来实现聚类任务等。

3. WASP: Scalable Bayes via barycenters of subset posteriors (AISTATS 2015)
使用离散形式的 2-W 距离，计算子数据集的后验分布的基于此距离的 barycenter 。

9. Gromov-Wasserstein Averaging of Kernel and Distance Matrices (ICML 2016) 🌟
定义了广义的熵正则化 G-W 距离。不同距离或核矩阵的质心 (barycenter) 等价于 Frechet 均值。

30. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes (NIPS 2017)
研究非退化的高斯过程，在 W-2 距离下可以定义唯一的 barycenter，并可以用有限维的表示进行高效运算。

STDA

2. Statistical Topological Data Analysis using Persistence Landscapes (JMLR 2015)
提出了一种更好的拓扑方法：持久性 landscape。证明了它可以给出 p-W 度量的下界。

PH

10. Subsampling Methods for Persistent Homology (ICML 2015)
使用子样本的方法简化 PH 的计算。对于一个大数据集，选取不同的子样本集，计算不同子样本集的 landscape，然后再综合起来。证明了平均的 landscape 在 W 度量的意义下是稳定的（对于其底部度量空间上的分布的扰动来说）。

GENERALIZATION OF PCA

[10. Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric (NIPS 2015)]

2-W。使用一种近似的投影梯度下降算法，计算出 PM 的泛化的测地线主成分，有可能对分析 PM 或图形有一定帮助。

[11. Principal Differences Analysis: Interpretable Characterization of Differences between Distributions (NIPS 2015)]

分析不同高维分布之间的差异。稀疏 PCA 实际上是 PDA 的一个特例。这里计算投影后的单变量 2-W 距离。

RELATIONSHIP TO CURVATURE

[12. Positive Curvature and Hamiltonian Monte Carlo (NIPS 2014)]

这里使用了 1-W 距离与 Ricci 曲率的关系，使用 1-W 定义了一种马尔可夫链的粗糙 Ricci 曲率。

[13. Empirical performance maximization for linear rank statistics (NIPS 2008)]

使用 ROC 曲率作为度量。文中提到了它与 1-W 距离的联系。

As a Metric

CONVERGENCE

[14. Parallelized Stochastic Gradient Descent (NIPS 2010)]

这里使用了 p-W 距离定义分布之间的距离。本文介绍了并行随机梯度下降的算法，并使用压缩映射方法量化了参数的分布到其渐进极限的收敛速度。

[4. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent (PMLR 2017)]

给出了 LMC 在 2-W 度量意义下的收敛性。本文讲解了 LMC 和 gradient descent 的联系。

[5. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Non-asymptotic Analysis (PMRL 2017)]

给出了 SGLD Markov 在 2-W 度量意义下的收敛性。

[6. Quantifying the accuracy of approximate diffusions and Markov chains (AISTATS 2017)]

使用了 1-W 定义两个扩散的平衡态分布的距离，用于给出界限。该方法在对 Langevin dynamics 的分析中也适用。

[12. Stochastic Bouncy Particle Sampler - Supplement (ICML 2017)]

使用 1-W 距离给出了偏差与真实分布之间的界限。使用了 Stein’s method。

[7. On the Ability of Neural Nets to Express Distributions (PMLR 2017)]

证明了任何紧支集的概率分布可以被有限层神经网络在 2-W 度量的意义下逼近。

OTHERS

15. Estimation of Intrinsic Dimensionality Using High-Rate Vector Quantization (NIPS 2005)

使用了 p-W 距离度量 quantizer mismatch，即度量噪声和原始 clean 数据的距离。

3. Ground Metric Learning (JMLR 2014)

本文使用了离散形式的最优传输距离。主要研究在 normalized histograms 间的距离的学习问题，即如何学习和选择最优的 ground metric。

13. Concentration in unbounded metric spaces and algorithmic stability (ICML 2014)

使用最优传输距离，得到样本 iid 猜想的松弛化结果。

5. Exploring and measuring non-linear correlations: Copulas, Lightspeed Transportation and Clustering (JMLR 2016)

这里使用熵正则化的离散最优传输距离，以减轻计算量。文中使用此种距离来度量非线性相关关系

14. Estimating the unseen from multiple populations (ICML 2017)

定义了 multiple populations 空间上的 earth-mover 度量。

Other Distances

16. The Perturbed Variation (NIPS 2012)

PV 距离可以视为是 W 距离和 TV (Total Variation) 距离之间的调和态。用来测定两组有限个数样本是否来自相似（不一定相同）的概率分布。

29. Population Matching Discrepancy and Applications in Deep Learning (NIPS 2017)

PMD 与 MMD 以及 W1 都有联系。N-PMD 的解是 W1 的强相合估计。

GAN

4. The Cramer Distance as a Solution to Biased Wasserstein Gradients (arXiv 2017)

提出了一种 Cramer 距离，结合了 W 距离和 KL 距离的优势。提出了 CGAN，并证明了它优于 WGAN。文中讲解了概率的 divergences 和 metrics ，提出 divergence 在机器学习任务中需要满足的三项特征：sum invariance, scale sensitivity, 和 unbiased sample gradients。其中 W 距离唯一不满足第三项，而 KL 距离满足，因此提出了 Cramer 距离来综合两种度量。

Convergence

LMC

18. Finite-Time Analysis of Projected Langevin Monte Carlo (NIPS 2015)

Projected LMC 是投影随机梯度下降的近亲，它可以视为是带有漂移的反射布朗运动。本文证明了通过选择合适的步长和循环次数，我们所研究的 projected LMC 的变量 $(\bar{X}_k)$ 服从的分布可以在 TV (Total Variation) 距离的意义下收敛到目标分布。本文利用了 1-W 与 TV 距离的关系（由 1-W 距离下的界限可以得到 TV 距离下的界限）得出这一结论。本文的成果可能对SGLD (Stochastic Gradient Langevin Dynamics) 算法等的研究有帮助。

Solving OT

17. Sinkhorn Distances: Lightspeed Computation of Optimal Transport (NIPS 2013)
Sinkhorn 距离，即是考虑熵正则化的 W 距离。文中提出此种情况可使用 Sinkhorn 算法，计算量较小。在图片分类任务中进行了测试。

8. Iterative Bregman Projections for Regularized Transportation Problems (SIAM.JSC 2015)
提供了最优传输中 linear programs 的近似解的数值框架。有关 barycenter 等变分问题的计算。

20. Stochastic Optimization for Large-scale Optimal Transport (NIPS 2016)
给出了求解大规模 OT 问题的一种算法。证明了这种算法优于 Sinkhorn 算法。在连续的问题中使用了 RKHS 上的随机优化。

21. Mapping Estimation for Discrete Optimal Transport (NIPS 2016)
通常的算法只是计算 OT 问题中的概率连接函数，本文提出了一种同时计算概率连接和近似的最优传输映射的方法。提出了这一方法与领域适应和图像编辑任务的联系。

20. A Simulated Annealing Based Inexact Oracle for Wasserstein Loss Minimization (ICML 2017)
本文基于模拟退火给出了（离散） WLM 问题的一种解法，以及使用 Gibbs 采样器去逼近 W losses 列的偏微分。比较了此种方法与熵正则化的方法。

21. Tensor Balancing on Statistical Manifold (ICML 2017)
与离散形式最优传输问题的计算相关。

26. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration (NIPS 2017)
结合 Sinkhorn 距离的计算。

Others

22. Basis refinement strategies for linear value function approximation in MDPs (NIPS 2015)
马尔可夫决策过程。建立了一个解决此类问题的普适框架。运用了 Kantorovich 度量，并涉及 Monge-Kantorovich 问题。

5. An Interpolating Distance between Optimal Transport and Fisher-Rao (arXiv 2015)
Twisting the geometry of the support (say, Wasserstein's optimal transport) with the geometry of the parametric distributions (Fisher-Rao geodesic distances).

6. Sub-modular Functions: from Discrete to Continuous Domains (arXiv 2016)
与 W 距离有一定对应关系。此类函数可用于计算 Robust Budget Allocation （属于最优分配问题）。见 Robust Budget Allocation via Continuous Sub-modular Functions (ICML 2017)。

23. Gradient descent GAN optimization is locally stable (NIPS 2017)

加入一个正则化项保证 WGAN 以及原始 GAN 优化的局部稳定。

27. The numerics of GANs (NIPS 2017)
新的训练算法。从能达到 Nash 均衡的角度，提出光滑的 2-player 问题，对广泛的度量（f-divergence， W 等）进行讨论。

28. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (NIPS 2017)
新的训练算法。适用于 WGAN。

Kernel Methods, IPM and RKHS

这部分内容实际上不属于最优传输的范畴，但与 1-W 有密切关联，而且与 GAN 相关很多最新进展相关。下面列出的主要是有提到 Wasserstein 的文章，主要集中在 2017年。

4. Hilbert Space Embeddings and Metrics on Probability Measures (JMLR 2010)🌟

详细介绍了一种拟度量，即概率积分度量 (IPM) 。IPM 具有较高的自由度，其性质取决于其测试函数（在 WGAN 中是判别器函数）在什么空间之中。1-W ，KL 等等都可视为其特殊情况。
本文重点讲解一种特别的 IPM，即基于 RKHS 嵌入的 IPM，测试函数限制在 RKHS 中的单位球中。讨论了此种度量优于其他 IPM 的性质，包括与 ϕ-divergence (f-divergence) 的对比、经验估计的收敛性、概率测度定义域的灵活性。可以说比较全面，目前很多 GAN 中用的度量都能在其中找到。

31. Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels (NIPS 2016)

提出 MMD 在有限样本估计中的下界。

Unsupervised Learning

15. Adversarial Feature Matching for Text Generation (ICML 2017)

实际上也是用一种核方法，使用了 RKHS 上的 Maximum Mean Discrepancy 度量。MMD 属于 IPM 。

GAN

16. McGan: Mean and Covariance Feature Matching GAN (ICML 2017)

基于 IPM 选择了一组特别的测试函数，使得该度量等于分布的 mean feature embedding 之间最远的 l-q 距离。它相应于分布之间的 MMD 度量。基于此度量提出了 McGAN 学习框架。

17. Generalization and Equilibrium in Generative Adversarial Nets (GANs) (ICML 2017), supplement

提出了 GANs 的泛化理论。定义了 “Generalize” 的概念，并证明了基于 JS 或 W 的 GAN 都不是 Generalize 的。定义了一种统领性的度量 F-distance （比 IPM 更甚），提出了泛化性更好的 Neural Net Distance。证明了在 GAN 的学习中使用混合的 G 和 D 会存在平衡态，基于此提出了 MIX+GANs 模型。

1. A Kernel Two-Sample Test (JMLR 2012), cited by this paper

24. Fisher GAN (NIPS 2017)

基于 IPM 框架。使用了 Fisher IPM。

25. MMD GAN: Towards Deeper Understanding of Moment Matching Network (NIPS 2017)

核对抗学习的方法结合 Generative moment matching network (GMMN)。基于 MMD 及其 weak-* 收敛的性质。

Convergence

19. Measuring Sample Quality with Stein's Method (NIPS 2015) ⭐

这里使用了 IPM 来判别样本的分布与目标分布间的距离，使用 Stein’s method 来帮助选择测试函数。

18. Measuring Sample Quality with Kernels (ICML 2017)

基于 IPM 以及 Stein’s method 定义了 kernel Stein discrepancy (KSD)， 用于分析 sample 到目标分布的收敛性。

19. Estimating individual treatment effect: generalization bounds and algorithms (ICML 2017)

使用 IPM 度量，给出了 W 距离下的以及 MMD 距离下的 ITE 预测的期望泛化误差 bounds。

Geometry / Topology

STDA

9. Statistical Topological Data Analysis - A Kernel Perspective (NIPS 2015)

利用代数拓扑中的方法可以研究数据的拓扑性质。持续同调（persistent homology）是用于在不同尺度下找到数据的“相关的”拓扑性质的工具。持久性图（persistence diagrams， PD）是该方法的输出，它编码不同性质的 life span。在 Statistical TDA 中常用 p-W 定义 PD 空间上的度量。但在机器学习任务中通常不研究这种复杂的空间，而是倾向于研究 Hilbert 空间中的学习。本文研究一种 STDA 的方法：将 PD 空间上的 PM 嵌入到再生核 Hilbert 空间（RKHS）中，并利用了 PD 空间上的一种核方法。

2. A Hilbert Space Embedding for Distributions (ALT 2007), cited by the above paper
3. Unifying Divergence Minimization and Statistical Inference via Convex Duality, cited by 2.

PD

11. Sliced Wasserstein Kernel for Persistence Diagrams (ICML 2017) ⭐

使用 Sliced W kernel 作为 1-W 距离的一种近似，定义了一种 PD 空间上新的 kernel。