*

	Γ	Δ	Θ	Λ	Ξ	Π	Σ	Υ	Φ	Ψ	Ω
\mit	$\mit\Gamma$	$\mit\Delta$	$\mit\Theta$	$\mit\Lambda$	$\mit\Xi$	$\mit\Pi$	$\mit\Sigma$	$\mit\Upsilon$	$\mit\Phi$	$\mit\Psi$	$\mit\Omega$

alpha	A	α	beta	B	β
gamma	Γ	γ	delta	Δ	δ
epsilon	E	ϵ, ε	zeta	Z	ζ
eta	H	η	theta	Θ	θ, ϑ
iota	I	ι	kappa	K	κ
lambda	Λ	λ	mu	M	μ
nu	N	ν	omicron	O	o
xi	Ξ	ξ	pi	Π	π, ϖ
rho	P	ρ, ϱ	sigma	Σ	σ, ς
tau	T	τ	upsilon	Υ	υ
phi	Φ	ϕ, φ	chi	X	χ
psi	Ψ	ψ	omega	Ω	ω

*: wangjunjie2013@gmail.com

1 An introduction to sampling via measure transport

Given a transport map, one can generate arbitrarilu many independent and unweighted samples from the target simply by pushig forward reference samples through the map.

Imagine generating samples x_i that are distributed according to u_{ref} and then applying T to each of these samples. Then the transformed samples $T(x_i)$ are distributed according to u_{tar}

- constructing transport maps given the ability to evaluate only the unnormalized probability density of the target
- constructing transport maps given only samples from a distribution of interest, but no explicit density.

1.1 Transport maps and optimal transport

Let $u_{tar} \in \mathcal{B}(R^n) \rightarrow R_+$ be a probability measure that we wish to characterize, defined over the Borel-σ-algebra on R^n. Let $u_{ref} \in \mathcal{B}(R^n) \rightarrow R_+$ be another probability measure from which we gen easily generate independent and unweighted samples,e.g. a standard Gaussian. Then a transport map $T : R^n \rightarrow R^n$ pushes forward u_{ref} to u_{tar} if and only if $u_{tar}(A) = u_{ref}(T^{-1}(A))$ for any set $A \in \mathcal{B}(R^n)$. We can write this compactly as:

$$T_\sharp u_{ref} = u_{tar} \tag{1}$$

A transport map T satifying (1) can be understood as a deterministic coupling of two probability measures.

There may be infinitely many such transormations, however. One way of choosing a particular map is to introduce a transport cost $c : R^n \times R^n \rightarrow R$ such that $c(x, z)$ represents the work needed to move a unit of mass from x to z. The resulting cost of a particular map is then

$$C(T) = \int c(x, T(x)) du_{ref}(x) \tag{2}$$

Minimizing (2) while simutaneously satisfying (1) corresponds to a problem first posed by Mone in 1781. The solution of this constrained minimization problem is the *optimal transport map*.

1.2 Direct transport: constructing maps from unnormalized densities

In this subsection we show how to construct a transport map that pushes forward a reference measure to the target measure when only evaluations of the *unnormalized target density* are available.

1.2.1 Preliminaries

We assume that both target and reference measures are absolutely continuous with respect to the Lebesgue measure on R^n. Let π and η be, respectively, the normalized target and reference densities with respect to the Lebesgue measure. We seek a diffeomorphism T(a smooth function with smooth inverse) that pushes forward the reference to the target measure,

$$u_{tar} = u_{ref} \circ T^{-1} \tag{3}$$

where $\circ$ denotes the composition of functions. In terms of densities, we will rewrite (3) as $\pi = T_\sharp \eta$. $T_\sharp \eta$ is the pushforward of the reference density under the map T, and it is defined as:

$$T_\sharp \eta := \eta \circ T^{-1} |det \nabla T^{-1}| \tag{4}$$

where ∇T^{-1} denotes the Jacobian of the inverse of the map.

if $(x_i)_i$ are independent samples from η, then $(T(x_i))_i$ are independnt samples from $T_\sharp \eta$. Hence, if we find a transport map T that satisfies $T_\sharp \eta = \pi$, then then $(T(x_i))_i$ will be independent samples from the target distribution. In particular, the change of variables formula:

$$\int g(x)\pi(x)dx = \int [g \circ T](x)\eta(x)dx \tag{5}$$

The map therefore allows for direct computation of posterior expectations.

1.2.2 Optimization problems

Let Y be an appropriate set of diffeomorhpisms. Then, any global minimizer of the optimization problem:

$$\begin{aligned} min \quad & D_{KL}(T_\sharp \eta || \pi) \\ s.t. \quad & det \nabla T > 0 \\ & T \in \mathrm{Y} \end{aligned} \tag{6}$$

In fact, any global minimizer of (6) achieves the minimum cost $D_{KL}(T_\sharp \eta || \pi) = 0$ and implies that $T_\sharp \eta = \pi$. The constraint det $\nabla T > 0$ ensures that the pushforward density $T_\sharp \eta$ is strictly postive on the support of the target.

AAmong these minimizers, a particularly useful map is given by the Knothe-Rosenblatt rearrangement.

Carlier, G., Galichon, A., Santambrogio, F.: From Knothes transport to Breniers map and a continuation method for optimal transport. SIAM Journal on Mathematical Analysis 41(6), 25542576 (2010)

We can furthe constrain (6) so that the Knothe-Rosenblatt rearrangement is the unique global minimizer of:

$$\begin{aligned} min \quad & D_{KL}(T_\sharp \eta || \pi) \\ s.t. \quad & det \nabla T \succ 0 \\ & T \in \mathrm{Y}_\Delta \end{aligned} \tag{7}$$

where Y_Δ is now the vector of smooth traingular maps. The constraint $det \nabla T \succ 0$ suffices to enforce invertibility of a feasible triangular map.

Let $\bar{\pi}$ denote any inormailized version of the target density. For any map T in the feasible set of (7), the object function can be written as:

$$D_{KL}(T_\sharp \eta || \pi) = D_{KL}(\eta || T_\sharp^{-1} \pi) = \mathbb{E}[-log \bar{\pi} \circ T - log det \nabla T] + \mathfrak{E} \tag{8}$$

$\mathfrak{E}$ is a term independent of the transport map and thus a constant for the purposes of optimization. The resulting optimization problem reads as:

$$\begin{aligned} min \quad & \mathbb{E}[-log \bar{\pi} \circ T - log det \nabla T] \\ s.t. \quad & det \nabla T \succ 0 \\ & T \in \mathrm{Y}_\Delta \end{aligned} \tag{9}$$

Notice that we can evaluate the objective of (9) given only the unnormalized density $\bar{\pi}$ and a way to compute the integral $\mathbb{E}_\eta[\cdot]$. There exist a host of techniques to approximate the integral with respect to the reference measure, including quadrature and cubature formulas, sparse quadratures, Monte Carlo methods, and quasi-Monte Carlo(QMC) methods.

(9) is a linearly constrained nonlinear differentiable optimization problem.

1.2.3 Convergence, bias, and approximate maps

A transport map provides a deterministic solution to the problem of sampling from a given unnormalized density, avoiding classical stochastic tools such as MCMC. A major concern in MCMC sampling methods is the lack of clear and generally ap-

plicable convergence criteria. In the transport map framework, on the other hand, the convergence criterion is borrowed directly from standard optimization theory.

The KL divergence $D_{KL}(T_\sharp\eta||\pi)$ can be estimated as

$$D_{KL}(T_\sharp\eta||\pi) \approx \frac{1}{2}\mathbb{V}ar_\eta[log\eta - logT_\sharp^{-1}\pi] \tag{10}$$

..... Too long.....

1.3 Inverse transport: constructing maps from samples

In this section, we assume that the target density is unknown and that we are only given a finite number of samples distributed according to the target measure. We show that under these hypotheses it is possible to efficiently compute an *inverse transport*– a transport map that pushes forward the target to the reference measure – via convex optimization.

1.3.1 Optimization problem

The inverse transport pushes forward the target to the reference measure:

$$\mu_{ref} = \mu_{tar} \circ S^{-1}$$

We focus on the inverse trangular transport because it can be computed via convex optimization given samples from the target distribution. It is easy to see that the monotone increasing Knothe-Rosenblatt ewarrangement that pushes forward μ_{tar} to μ_{ref} is the unique minimizer of

$$\begin{aligned} min \quad & D_{KL}(T_\sharp\eta||\pi) \\ s.t. \quad & det\nabla S \succ 0 \\ & S \in \mathcal{T}_\Delta \end{aligned} \tag{11}$$

For any map S in the feasible set of (11), the objective function can be written as:

$$\begin{aligned} D_{KL}(S_\sharp\eta||\pi) &= D_{KL}(\pi||S_\sharp^{-1}\eta) \\ &= \mathbb{E}_\pi[-log\eta \circ S - logdet\nabla S] + \mathfrak{e} \end{aligned} \tag{12}$$

where $\mathfrak{e}$ is a term independent of the transport map and thus a constant for the purposes of optimization. The resulting optimization problem is a stochastic program given by

$$\begin{aligned} \min \quad & \mathbb{E}_\pi[-log\eta \circ S - logdet\nabla S] \\ s.t. \quad & det\nabla S \succ 0 \\ & S \in \mathcal{T}_\Delta \end{aligned} \tag{13}$$

A sample-average approximation(SAA) of (13) is given by:

$$\begin{aligned} \min \quad & \frac{1}{M}\sum_{i=1}^{M}[-log\eta(S(z_i)) - logdet\nabla S(z_i)] \\ & s.t. \quad \partial_k S^k \succ 0 \\ & S \in \mathcal{T}_\Delta \end{aligned} \tag{14}$$

1.3.2 Convexity and separability of the optimization problem

Let the reference measure be standard Gaussian. In this case, (14) can be written as

$$\begin{aligned} \min \quad & \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{n}\left[\frac{1}{2}(S^k)^2(z_i) - log\partial_k S^k(z_i)\right] \\ & s.t. \quad \partial_k S^k \succ 0 \\ & S \in \mathcal{T}_\Delta \end{aligned} \tag{15}$$

All the components of the inverse transport can be computed independently and in parallel by solving optimization problems.

1.3.3 Computing the inverse map

We recommend running a few Newton iterations of the form

$$z_{j+1} = z_j - \nabla S(z_j)^{-1}(S(z_j) - x*)$$

An alternative way to evaluate the direct transport is to build a parametric representation of T itself via standard regression techniques. In particular, if $\{z_1, ...z_M\}$ are samples from the target distribution, then $\{x_1, ..., x_M\}$, with $x_k := S(z_k)$ for $k = 1, ..., M$, are samples from the reference distribution. We can use these pairs of samples to define a simple constrained least-squares problem to approximate the direct transport as:

$$\begin{aligned} \min \quad & \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{n}(T^i(x_k) - z_k)^2 \\ & s.t. \quad \partial_k T^k \succ 0 \\ & T \in \mathcal{T}_\Delta \end{aligned} \tag{16}$$

1.4 Parameterization of transport maps

We must define finite-dimensional approximation spaces(e.g. $\mathcal{T}_\Delta^h$) within which we search for a best map.

1.4.1 Polynomial representations

A natural way to parameterize each component of the map T is by expanding it in a basis of multivariate polynomials. The univariate polynomials can be chosen from any standard orthogonal polynomial family(e.g., Hermite, Legendre, Laguerre) or they can even be monomials.

I use the scipy.special package to implement the hermit. The example code is as follows:

Hermit Polynomial univariate:

$$H_k = (-1)^k e^{x^2/2} d_x^k e^{-x^2/2}$$

```
import numpy as np
from scipy.special import eval_hermitenorm

class Hermite(object):
    """docstring for Hermite"""
    def __init__(self):
        return

    def hermite_value(self,x,order):
        '''
            return the hermite polynomial value
        '''
        return eval_hermitenorm(order,x)
    def grad_value(self,x,order):
        '''
            return the first derivate of hermite polynomial evaluated at
                                                    x
        '''
        return order*eval_hermitenorm(order-1,x)


if __name__ == '__main__':
    '''
        test
    '''
    import numpy as np
    import matplotlib.pyplot as plt
    hermite = Hermite()
    x = np.linspace(-5,5)
    fig,axes = plt.subplots(5,1)
    for i in xrange(0,5):
```

```
        y = hermite.hermite_value(x,i)
        axes[i].plot(x,y,label = "order{0}".format((i)))
    plt.show()



    plt.legend(loc="best")
    plt.savefig('figs/hermite_poly.pdf')
```

Use multivariate polynomials, we can express each component of the transport map as

$$T^k(x) = \sum_{j \in \mathcal{J}_k} \gamma_{k,j} \psi_j(x), k = 1, ..., n \tag{17}$$

where $\mathcal{J}$ is a set of multi-indices defining the polynomial terms in the expansion for dimension k and γ is a scalar coefficient.

1.4.2 Radial basis functions

An alternative to a polynomial parameterization of the map is to employ a combination of linear term and radar basis functions. This representation can be more efficient than a polynomail representation in certain cases.

2 Inference via low-dimensional couplings

The transport map T can be viewed as a transormation that moves particles : given a collection of samples from v_η , T rearranges them in accordance with the new distribution v_π

Optimal transprt maps, for instance, define couplings that minimize a particular integrated transport cost expressing the effort required to rearrage samples. In recent years, several other couplings have been proposed for use in statistical problems, e.g.,

- parametric approximations – Moselhy, T. and Marzouk, Y. (2012). Bayesian inference with optimal maps. Journal of Computational Physics 231 78157850.
- Knote-Rosenblatt rearrangement– Rosenblatt, M. (1952). Remarks on a multivariate transformation. The Annals of Mathematical Statistics 470472
- coupling induced by ODE flows— Heng, J., Doucet, A. and Pokern, Y. (2015). Gibbs flow for approximate transport with applications to Bayesian computation. arXiv:1509.08787. Daum, F. and Huang, J. (2008). Particle flow for nonlinear filters

with log-homotopy. In SPIE Defense and Security Symposium 696918696918. International Society for Optics and Photonics. Anderes, E. and Coram, M. (2012). A general spline representation for nonparametric and semiparametric density estimates using diffeomorphisms. arXiv:1205.5314.

Yet the construction, representation, and evaluation of all these maps grows challenging in high dimensions.

The central contribution of this paper is to **establish a link between the conditional independence structure of the target measure and the existence of special low dimensional coupling. These couplings are induced by transport maps that are sparse or decomposable.**

- sparse: A sparse map consists of scalar-valued compnent functions that each depend only a few input variables
- decomposable map: A decomposable map factorizes as the exact composition of finitely many functions of low effective dimension(i.e, $T = T_1 \circ ... \circ T_l$, where each T_i differs from the identity map only along a subset of its components).

The utility of thses results is twofold:

- First, they make the construction of couplings tractable for a large class of inference problems.
- Second, they suggest new algorithmic approaches for important classed of statistical models.

2.1 Notation

- $f \circ g$: the composition of f, g
- $\partial_k f$ partial derivative of f with respect to its kth input variable
- $x \to Qx$: linear map
- $T_\sharp v$: pushforward measure given by $v \circ T^{-1}$
- $T^{-1}(B)$: set-valued preimage of B under T
- $T^\sharp v$: the pullback measure given by $v \circ T$
- $T_\sharp \pi$:

2.2 Triangular transport maps: a building block

3 An Optimal Transport Formulation of the Linear Feedback Particle Filter

Based on the concept of optimal transportation, a time-stepping optimization procedure is proposed here to obtain a unique optimal control law, denoted as μ_t^* and K_t^*. In this procedure, a finite time interval is divided into discrete time steps $t_0, ..., t_n$. Then a discrete time random process, S_{t_k} is constructed by initializing S_{t_0} according to the initial prior $P(X_0)$, and sequentially evolving $S_{t_k} \to S_{t_{k+1}}$ at each time-step with a map denoted by T_k:

$$S_{t_{k+1}} = T_k(S_{t_k}), S_0 P(X_0) \tag{18}$$

3.1 Exactness and non-uniqueness

4 transportmaps.mit.edu

4.1 Bayesian models

Bayesian models arise naturally in statistical inference problems, where the belief represented by a prior probability distribution need to be updated according to some observations. We can summarize these models as follows.

Let $G : \mathbb{R}^d \to \mathbb{R}^{d_z}$ be a map, let $X \in \mathbb{R}^d$ be a random variable with distribution v_ρ, representing the prior belief on the state of X. Let $d \in \mathbb{R}^{d_y}$ be observations of the output of G for some unknown inputs $x \in \mathbb{R}^d$ obtained through the observations. This model for the measurements

$$d = f(G(x), \xi) \tag{19}$$

we defin the posterior distribution v_π by its density

$$\pi(x|Y = d) \approx \mu(f^{-1}(G(x), d))\rho(x) = \mathcal{L}(x)\rho(x)$$

$\mathcal{L}(x)$ is the likelihood and $\rho(x)$ is the prior density.

4.1.1 Linear Gaussian model

Let us consider here the linear model $G(x) = Gx$ with additive Gaussian noise

$$d = Gx + \xi$$

This means that the likelihood is defined as

$$\mathcal{L}_d(x) = \mu(d - Gx)$$

Let the prior distribution on X be Gaussian, $v_\rho \sim \mathcal{N}(m_\rho, \Sigma_\rho)$.

Feedback particle filter: The feedback particle filter for linear Gaussian problem is given by

4.2 Background on optimal transportation

Let P_X and P_Y be two given probability measures on $\mathbb{R}^d$ with finite second moments. The optimal transportation problem is to minimize

$$min_T \quad E[(T(x) - x)^2]$$

over all maps $T : \mathbb{R}^d \to \mathbb{R}^d$ such that:

$$X \sim P_X, \quad T(x) \sim P_Y$$

If it exists, the minimizer T^*, is called the optimal transport map between P_X and P_Y. The optimal cost is referred to as L^2-Wasserstein distance between P_X and P_Y.

4.2.1 Optimal map between Gaussians

Suppose P_X and P_Y are Gaussian distributions, $\mathcal{N}(X, \Sigma_X)$ and $\mathcal{N}(Y, \Sigma_Y)$. Then the optimal transport map between P_X and P_Y is given by:

$$T(x) = Y + F(x - X)$$

where

$$F = \Sigma_Y^{\frac{1}{2}}(\Sigma_Y^{\frac{1}{2}}\Sigma_X\Sigma_Y^{\frac{1}{2}})^{-\frac{1}{2}}\Sigma_Y^{\frac{1}{2}}$$

5 Classical Monge-Kantorovich problem

In our simplest exercise we generate two of these random histograms and then try to design a transport plan between them. The concise statement of the problem is

$$min\left\{TrM^TZ|Z1 = p, Z^T1 = q\right\}$$

where p, q are m vectors representing the mass of the histogram bars for the source and destination histograms, respectively. The matrix M represents the distances between

the pairs of historgram bars, i.e. $M_{i,j} = |x_i - y_j|$. The trace formulation of the objective is simply the discrete analogue of the continuous formulation

$$min_{\pi \in \prod(\mu,v)} \int c(x,y) d\pi(x,y)$$

where $\pi(\mu, v)$ for all borel sets A and B. The matrix Z in the discretized problem playes the role of π in the continuous formulation, indicating how much mass is to be transfered from points x to point y.

6 Optimal Transport Filtering with Particle Reweighing in Finance

Particle flow filter(daum) allows the reduction of the number of particles we need in order to get a tolerable level of errors in the filtering problem. The main idea behind this method is the evolution in homotopy parameter λ from prior to the target density. The introduced a particle flow, in which particles are gradually transported without the necessity to randomly sample from any distribution.

The idea of transportation and rewighing mechanism is to transport particles through the sequence of densities that move the least during the synthetic time until they reach the posterior distribution. By regenerating particles according to their weight at each time step we are able to direct the flow and further minimize the variance of the estimates.