Mathematical Details

This library is entirely focused on implementing the *transport map accelerated Markov chain Monte Carlo* inference algorithm discussed in the following paper.

> Parno, Matthew D., and Youssef M., Marzouk. "Transport Map Accelerated Markov Chain Monte Carlo". *SIAM/ASA Journal on Uncertainty Quantification* 6, no.2 (2018): 645-682. SIAM Link, arXiv Link.

In summary, this is an inference algorithm which samples a posterior distribution `\mu_\theta` by simultaneously finding a map `T` which transports said distribution to a *reference* standard Gaussian `\mu_r = T_\# \mu_\theta` and performing Metropolis-Hastings in this reference space. The proceeding subsections will describe the objects at play.

Transport maps

Suppose we have some distribution `\mu_\theta` with density `\pi` with respect to the Lebesgue measure on `\mathbb R^d`.

```
\displaystyle \mu_\theta({\rm d}\theta) = \pi(\theta) {\rm d}\theta
```

Provided a measurable map `T: \mathbb R^d \rightarrow \mathbb R^d`, we may *transport* `\mu_\theta` to another distribution `T_\#\mu_\theta`, defined to act on measurable functions `f: \mathbb R^d \rightarrow \mathbb R` like so.

```
\displaystyle \int f(r) T_\#\mu_\theta({\rm d}r) = \int f\big(T(\theta)\big) \mu_\theta({\rm
```

!!! note "Note" From a Monte Carlo perspective, a sample `r \sim T_\#\mu_\theta` is equivalent to sampling `\theta \sim \mu_\theta` and taking `r = T(\theta)` (hence the phrase *Transport*).

If our density `\pi` is continuous and `T` is a continuously differentiable bijection, the change of variables theorem from calculus tells us that the transport distribution (let's denote this `\mu_r = T_\#\mu_\theta`) will have the following density `p`.

```
p(r) = \pi\big(T^{-1}(r)\big) \big|\det \nabla T^{-1}(r)\big| \\[1em]
\displaystyle\int f(r) \mu_r({\rm d}r) = \int f(r) \pi\big(T^{-1}(r)\big) \big| \det \nabla
```

Above and throughout, `\nabla G` denotes the Jacobian matrix of a map `G: \mathbb R^d \rightarrow \mathbb R^d`. From the perspective of the Gen ecosystem, if an address `:theta` is intended to encode samples of `\mu_\theta`, then one can subsequently encode samples of `\mu_r = T_\ast\mu_\theta` to an address `:r` with the Trace Transform DSL.

The paper above concerns itself with finding a transport map (referred to as the *Knothe-Rosenblatt rearrangement*) `T: \mathbb R^d \rightarrow \mathbb R^d` such that `\nabla T` is lower-triangular and `\mu_r = T_\#\mu_\theta` is the standard Gaussian measure on `\mathbb R^d`. The lower-triangular property of `\nabla T` is equivalent to `T` having the following structure.

$$T(\theta_1, \ldots, \theta_d) = \Big(T_1(\theta_1), T_2(\theta_1, \theta_2), \ldots, T_d(\t$$

Such a structure is computationally advantageous, as the inverse image $T^{-1}(r)$ and Jacobian determinant $\det \nabla T(\theta)$ are easy to evaluate. Also, the Gaussian nature of $\mu_r = T_\#\mu_\theta$ means that sampling $\theta \sim \mu_\theta$ is as easy as sampling from standard Gaussian $r \sim \mu_r$ and evaluating $\theta = T^{-1}(r)$. Hence, having such a map T, or a nice approximation thereof, means we may efficiently sample complicated posterior distributions μ_θ; such an algorithm is apt for systems like Gen.

Approximating transport maps

In practice, it is infeasible to actually get the transport constraint $T_\#\mu_\theta = \mu_r$ for a standard Gaussian μ_r. Thus, for a fixed distribution μ_θ and proposed transport map $\tilde T$, it is imperative to measure the discrepancy of our transport from the standard Gaussian. In other words, we want to measure the effectiveness of the following approximation.

$$\tilde T_\#\mu_\theta \approx \mu_r$$

One solution to this is to recognize that the true constraint $\tilde T_\#\mu_\theta = \mu_r$ is equivalent to $\tilde T^{-1}_\#\mu_r = \mu_\theta$; this way, we may consider the discrepancy of an equivalent approximation.

$$\mu_\theta \approx \tilde T_\#^{-1}\mu_r$$

Denoting $\tilde\pi$ as the density of $\tilde T_\#^{-1}\mu_r$, we have

$$\tilde\pi(\theta) = p\big(\tilde T(\theta)\big) \big|\det\nabla\tilde T(\theta)\big|.$$

From here, we may measure the discrepancy between μ_θ and $\tilde T_\#^{-1}\mu_r$ with the Kullback-Leibler divergence.

$$\begin{aligned}
  D_{KL}\big(\mu_\theta \parallel \tilde T_\#^{-1}\mu_r \big)
  &= \int \log\Big(\frac{\tilde\pi(\theta)}{\pi(\theta)}\Big) \mu_\theta({\rm d}\theta) \\
  &= \int \log\pi(\theta) \mu_\theta({\rm d}\theta) + \int \Big(- \log p\big(\tilde T(\thet
\end{aligned}$$

This divergence is minimized when our transport constraint is exact, and so finding the true transport T is equivalent to solving the following optimization problem over the set $\mathcal T$ of lower-triangular continuously differentiable bijections.

$$T = \argmin_{\tilde T \in \mathcal T} \int \Big(- \log p\big(\tilde T(\theta)\big) - \log\b$$

If provided samples $\theta^{(1)}, \ldots, \theta^{(K)}$ from μ_θ, we may approximate the integral above as follows.

```
\begin{aligned}
  &\int \Big( - \log p\big(\tilde T(\theta)\big) - \log\big|\det\nabla\tilde T(\theta)\big|
  &\quad\approx K^{-1}\sum_{k=1}^K \Big( -\log p\big(\tilde T(\theta^{(k)})\big) - \log\big
  &\quad= K^{-1}\sum_{k=1}^K \Big( \frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^d \tilde T_
  &\quad= \frac{n}{2}\log(2\pi) + K^{-1} \sum_{i=1}^d \sum_{k=1}^K \Big( \frac{1}{2}\tilde T
\end{aligned}
```

Note that the first equality above is utilizing the closed form of the standard Gaussian density `p` and the lower-triangular structure of `\nabla\tilde T(\theta)`. With this approximation, we choose to instead minimize the following objective function.

```
C(\tilde T) = \sum_{i=1}^d\sum_{k=1}^K \Big(\frac{1}{2} \tilde T_i^2(\theta) - \log\frac{\pa
```

From here, we may reduce the optimization problem from the large space `\mathcal T` to a parameterized set of maps `\{T(\cdot; \gamma)\}_{\gamma \in \mathcal A}` where `\mathcal A` is some nice parameter set. In particular, for each `i=1,\ldots, d`, we may pick a finite ordered basis of functions `(\psi_j)_{j\in\mathcal J_i}` and declare our map parameterization so that our parameter set is `\mathcal A = \prod_{i=1}^d \mathbb R^{\mathcal J_i}` and the components of the map `T(\cdot; \gamma)` are as follows.

```
\tilde T_i(\theta; \gamma_i) = \sum_{j\in\mathcal J_i} \gamma_{i,j} \psi_j(\theta)
```

This linear form makes optimizing `C\big(\tilde T(\cdot;\gamma)\big)` simpler.

Map-based Markov chain Monte Carlo

Provided a complicated posterior distribution `\mu_\theta` and a transport map `\tilde T` such that `\tilde T_\#\mu_\theta` is approximately Guassian, we may choose to perform the Metropolis-Hastings algorithm on either of `\mu_\theta` or `\tilde T_\#\mu_\theta` and subsequently map samples via `\tilde T`. Because `\tilde T_\#\mu_\theta` is approximately Gaussian, our proposals do not need to account for complicated features of the distribution, as they would for `\mu_\theta`. To this end, the map `\tilde T` is accounting for these complicated features.

For a fixed proposal kernel `Q_r` with density `q_r`,

```
Q_r({\rm d}r | r') = q_r(r | r') {\rm d}r,
```

the subsequent pullback kernel designed to make the diagram commute