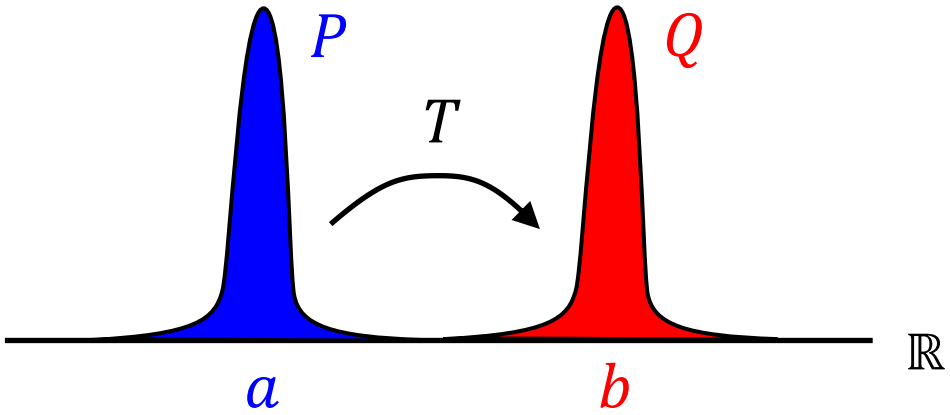
Finite sample rates for optimal transport estimation problems

Jan-Christian (JC) Hütter

jchuetter.web@gmail.com

Optimal transport and Wasserstein distance



$$D(P \,\|\, Q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx = \infty$$

$$TV(P, Q) = \frac{1}{2}\int |p(x) - q(x)|\, dx = 1$$

$W_2^2(P, Q) = \|a - b\|_2^2$ (Wasserstein distance)

- Common statistical distances insensitive to geometry of underlying space
- Monge problem: For P, Q probability distributions on $\mathbb{R}^d$, search for *transport map* $T_0 \colon \mathbb{R}^d \to \mathbb{R}^d$ that solves

$$W_2^2(P, Q) = \min\left\{\int \|T(x) - x\|_2^2\, dP(x) : T_{\#}P = Q\right\}, \qquad \text{(Monge)}$$

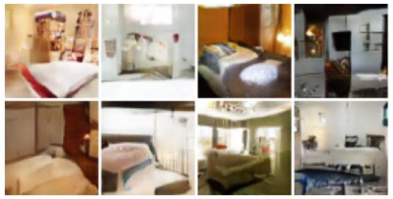
  where $T_{\#}P$ is push-forward of P, distribution of $Y = T(X),\ X \sim P$.
- Focus on square Euclidean cost $\|T(x) - x\|_2^2$, but other cost functions also of interest

Applications

Distance between probability measures (W_2)

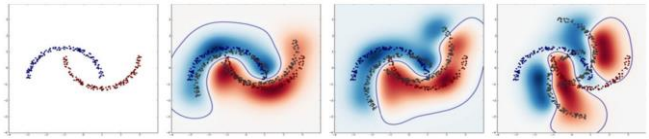
Bag-of-words models
(Rolet, Cuturi, Peyré, 2016)

Multi-label prediction
(Frogner et al., 2015)

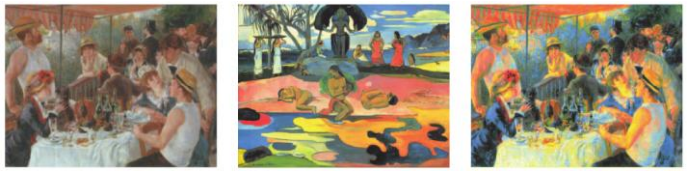


Wasserstein GAN
(Arjovsky, Chintala, Bottou, 2017)

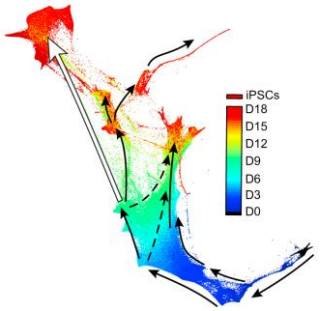
Uncoupled function estimation (T_0)



Domain adaptation
(Courty, Flamary, Tuia, 2017)

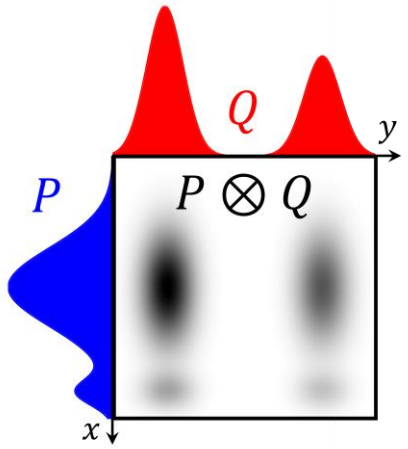


Color transfer
(Rabin, Delon, Gousseau, 2010)

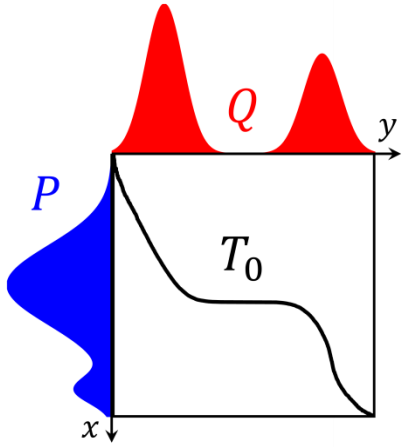


Trajectory inference in scRNA-Seq
(Schiebinger, Shu, Tabaka, et al., 2019)

Relax Monge problem to Kantorovich problem



Independent



Deterministic

- Reminder:

$$\min_{T_{\#}P=Q} \int \|T(x) - x\|_2^2 \, dP(x) \qquad \text{(Monge)}$$

- Constraint $T_{\#}P = Q$ highly non-linear, asymmetric
- Relaxation: Look for a transport plan γ_0 that solves

$$\min_{\gamma \in \Gamma(P,Q)} \int \|y - x\|_2^2 \, d\gamma(x,y), \qquad \text{(Kantorovich)}$$

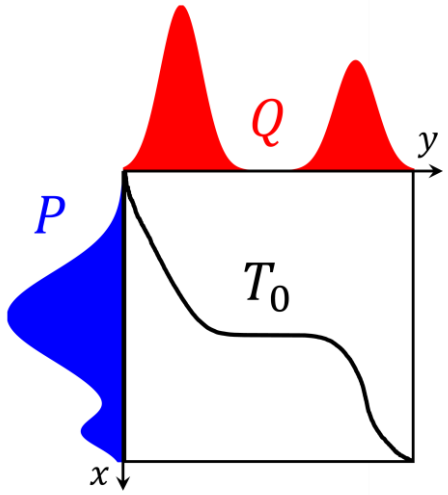
where $\Gamma(P,Q)$ is the set of probability measures on $\mathbb{R}^{d \times d}$ with marginals P and Q,

$$\int d\gamma(.,y) = P\,, \quad \int d\gamma(x,.) = Q$$

Brenier's theorem links Monge and Kantorovich problems

$$T_0 \in \arg\min_{T_\# P = Q} \int \|T(x) - x\|_2^2 \, dP(x) \qquad \text{(Monge)}$$

$$\gamma_0 \in \arg\min_{\gamma \in \Gamma(P,Q)} \int \|y - x\|_2^2 \, d\gamma(x,y) \qquad \text{(Kantorovich)}$$

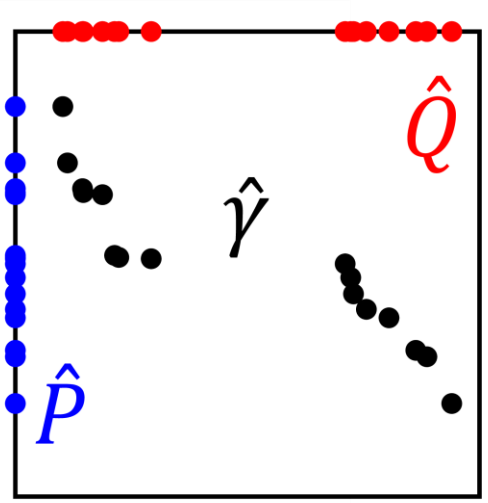


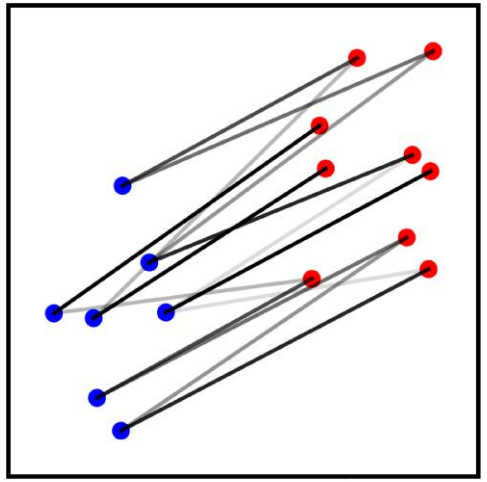
Theorem (Brenier)

- If P absolutely continuous measure, then γ_0 is unique and concentrated on graph of function T_0 that solves (Monge).
  Moreover, $T_0 = \nabla f_0$, the gradient of a convex function $f_0 : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$.
- Conversely, given P and $T_0 = \nabla f_0(x)$ for f_0 convex, T_0 solves (Monge) for P and $Q = (T_0)_\# P$.

Statistical optimal transport

Statistical optimal transport





- Sample access: observe $X_1, \ldots, X_n \sim P$, $Y_1, \ldots, Y_m \sim Q$ i.i.d.
- Empirical distributions

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \hat{Q} = \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}$$

- Can directly compute on empirical distributions

$$\widehat{W}_2^2 = W_2^2(\hat{P}, \hat{Q}) = \min_{\gamma \in \Gamma(\hat{P}, \hat{Q})} \int \|y - x\|_2^2 \, \mathrm{d}\gamma(x, y)$$

- Fast algorithms available: linear programming, entropic regularization
- Is this a good idea?
  - No out-of-sample prediction, but could do barycentric projection and 1-nearest-neighbor
  - Statistical guarantees?

Three related estimation problems

1. How well does $\hat{P}$ approximate P in W_2:

$$W_2^2(P, \hat{P}) \lesssim \ ?$$

2. How well can I estimate $W_2(P, Q)$:

$$\left|\widehat{W}_2^2(P, Q) - W_2^2(P, Q)\right| \lesssim \ ?$$
$$\left|\widehat{W}_2(P, Q) - W_2(P, Q)\right| \lesssim \ ?$$

By triangle inequality, $\left|W_2(\hat{P}, \hat{Q}) - W_2(P, Q)\right| \leq W_2(\hat{P}, P) + W_2(\hat{Q}, Q)$, so can get 2 from 1 for plug-in estimator

3. How well can I estimate T_0 or γ_0? For example,

$$\int \left\|\hat{T}(x) - T_0(x)\right\|_2^2 dP(x) \lesssim \ ?$$

Downside of geometry: curse of dimensionality

Theorem (Dudley, 1969; Weed, Bach, 2017)

$P \sim \mathrm{Unif}\left([0,1]^d\right)$, $X_1, \ldots, X_n \sim P$ i.i.d., then

$$W_2^2\left(P, \hat{P}\right) \gtrsim n^{-2/d} \quad \text{almost surely.}$$

Proof.

- Set $\epsilon = \frac{1}{100} n^{-1/d}$. By a volume argument, $X_i + \epsilon[-1,1]^d$ does not cover more than half of $[0,1]^d$. Denote that half by A
- A needs to get transported to $\hat{P}$, so each point travels at least ϵ.
- Hence, for any coupling γ,

$$W_2^2\left(P, \hat{P}\right) \geq \int \|y - x\|_2^2 \, d\gamma(x, y) \geq \frac{1}{2}\epsilon^2 = \frac{1}{2(100)^2} n^{-\frac{2}{d}}$$

□

Dimensionality-dependent rates

For the rest of the talk, assume P, Q compactly supported, $d \geq 5$.

Theorem (Dudley, 1969; Fournier, Guillin, 2014; Weed, Bach, 2017)
If P absolutely continuous, then

$$\mathbb{E}\left[W_2^2\left(P, \hat{P}\right)\right] \asymp n^{-\frac{2}{d}}.$$

- Also holds true when P supported on k-dimensional compact manifold (Weed, Bach, 2017):

$$\mathbb{E}\left[W_2^2\left(P, \hat{P}\right)\right] \asymp n^{-\frac{2}{k}}.$$

- Two-sample case (Chizat et al., 2020; Manole, Niles-Weed, 2021):

$$\mathbb{E}\left[\left|W_2^2\left(\hat{P}, \hat{Q}\right) - W_2^2(P, Q)\right|\right] \asymp n^{-\frac{2}{d}}$$

- Corollary

$$\mathbb{E}\left[\left|W_2\left(\hat{P}, \hat{Q}\right) - W_2(P, Q)\right|\right] \lesssim n^{-\frac{1}{d}} \wedge \frac{1}{W_2(P, Q)} n^{-\frac{2}{d}}$$

- Lower bounds (under some assumptions, Niles-Weed, Rigollet, 2020):

$$\mathbb{E}\left[\left|\widehat{W}_2^2 - W_2^2(P, Q)\right|\right] \gtrsim (n \log n)^{-\frac{2}{d}}$$

Lower dimensional transport maps

- What if distributions are full dimensional, but transport is not?
- Kolouri et al, 2019; Paty, Cuturi, 2019; Niles-Weed, Rigollet, 2019; Lin et al., 2021:

$$PW_{2,k}^2(P,Q) = \sup\left\{W_2^2\left((\Pi_E)_\# P, (\Pi_E)_\# Q\right) : E \; k-\text{dim subspace}\right\},$$

  where Π_E denotes the projection onto subspace E
- *Projection Robust Wasserstein distance/Wasserstein projection pursuit*
- Could also average instead of taking supremum, 1D case then known as sliced Wasserstein distance
- Recover lower-dimensional rates

$$\mathbb{E}\left[PW_{2,k}^2\left(\hat{P}, P\right)\right] \lesssim n^{-\frac{2}{k}}$$

- Computationally challenging, but amenable to Riemannian optimization (Lin et al., 2020)

Non-parametric smoothness regularization

Smooth densities

- Empirical distribution $\hat{P}$ is poor approximation to P
- For $\mathrm{M} > 0, \alpha > 1$, assume
  - P admits $C^{\alpha-1}$ smooth densities (also denoted by P)
  - Densities upper and lower bounded, $M^{-1} \leq P(x) \leq M$
  - Some additional assumptions on support
- Niles-Weed, Berthet, 2019: For wavelet density estimators $\tilde{P}$,

$$\mathbb{E}\left[W_2^2\left(P, \tilde{P}\right)\right] \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}}$$

- Under additional assumptions, can obtain similar results for $W_2^2(\tilde{P}, \tilde{Q})$ and $\int \left\|\tilde{T}(x) - x\right\|_2^2 dP(x)$ with $\tilde{T}$ transport map from $\tilde{P}$ to $\tilde{Q}$ (Manole et al., 2021; Deb et al., 2021)
- Shortcomings: to compute Wasserstein distance or transport map, still need to sample a lot!

Smooth transport maps

Theorem (H., Rigollet, 2021)

If

- $T_0 = \nabla f_0$ with $\mathrm{M}^{-1}\mathrm{I} \preccurlyeq \nabla^2 f_0 \preccurlyeq MI$ (f_0 strongly convex, i.e., $f_0 - M^{-1}\|\,.\,\|_2^2$ convex)
- T_0 is C^α-smooth on hypercube containing $\mathrm{supp}(P)$, $\alpha > 1$,
- $Q = (T_0)_\# P$

Then, there exists $\hat{T}$ such that

$$\mathbb{E}\int \left\|\hat{T}(x) - T_0(x)\right\|_2^2 dP(x) \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}} \log(n)^2$$

- Minimax optimal up to logs for classes mimicking the above assumptions

Exponent $2\alpha/(2\alpha - 2 + d)$

- Change of variables:

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(y) \asymp \frac{1}{\det \partial T\big(T^{-1}(y)\big)}, \quad y \in T(\mathrm{supp}(P))$$

- KL between "signals" is governed by ∂T, while estimation target is T
- Compare white noise model: estimation rate $n^{-2(\beta-k)/(2\beta+d)}$ for estimation of kth derivative of C^β function
- Set $\beta = \alpha - 1$, $k = -1$, estimation of anti-derivative of $C^{\alpha-1}$ function

Caffarelli's Regularity Theory: from smooth densities to smooth transport maps

- $P \in C^{\alpha-1}(\Omega_P), Q \in C^{\alpha-1}(\Omega_Q)$ for $\alpha > 1$
- P, Q upper and lower bounded on Ω_P, Ω_Q
- Ω_P, Ω_Q convex with C^2-boundary and uniformly convex

Then, $T_0 \in C^{\alpha}$

Simple upper bounds: empirical risk minimization

- For a moment, think about regression problem with data $\{(X_i, Y_i)\}_{i=1,\dots,n}$ and regression function T instead
- Risk function $\mathcal{S}(T) = \mathbb{E}\left[\|T(X) - Y\|_2^2\right]$
- Empirical risk: $\hat{\mathcal{S}}(T) = \frac{1}{n}\sum_{i=1}^{n}\|T(X_i) - Y_i\|_2^2$

$$T_0 = \arg\min_T \mathcal{S}(T), \qquad \hat{T} = \arg\min_T \hat{\mathcal{S}}(T)$$

- Bound excess risk

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{T}(X) - T_0(X)\right\|_2^2\right] \lesssim\ & \mathcal{S}(\hat{T}) - \mathcal{S}(T_0) = \mathcal{S}(\hat{T}) - \hat{\mathcal{S}}(\hat{T}) + \underbrace{\hat{\mathcal{S}}(\hat{T}) - \hat{\mathcal{S}}(T_0)}_{\leq 0} + \hat{\mathcal{S}}(T_0) - \mathcal{S}(T_0) \\ & \leq \mathcal{S}(\hat{T}) - \hat{\mathcal{S}}(\hat{T}) + \hat{\mathcal{S}}(T_0) - \mathcal{S}(T_0) \\ & \leq 2\sup_T |\mathcal{S}(T) - \hat{\mathcal{S}}(T)|\end{aligned}$$

- Control via empirical process theory, e.g. Dudley integral and covering numbers based on complexity of function class T
- Stability/margin condition

Proof strategy

- Find suitable risk function to apply empirical process theory
- Rewrite objective: expanding squares yields

$$\min_{\gamma \in \Gamma(P,Q)} \int \|y - x\|_2^2 \, \mathrm{d}\gamma(x,y) \Leftrightarrow \max_{\gamma \in \Gamma(P,Q)} \int \langle x, y \rangle \, \mathrm{d}\gamma(x,y)$$

Dual and semi-dual

Primal

$$\gamma_0 \in \arg\max \int \langle x, y \rangle \, \mathrm{d}\gamma(x, y)$$

$$\text{s.t. } \gamma \in \Gamma(P, Q)$$

Dual

$$(f_0, g_0) \in \arg\min \int f(x) \, \mathrm{d}P(x) + \int g(y) \, \mathrm{d}Q(y)$$

$$\text{s.t. } f(x) + g(y) \geq \langle x, y \rangle, \quad x, y \in \mathbb{R}^d$$

- Strong duality: for optimal γ_0, f_0, g_0,

$$\int \langle x, y \rangle \, \mathrm{d}\gamma_0(x, y) = \int \big(f_0(x) + g_0(y)\big) \, \mathrm{d}\gamma_0(x, y) = \int f_0(x) \, \mathrm{d}P(x) + \int g_0(y) \, \mathrm{d}Q(y)$$

- Semi-dual: Optimizing g for given f,

$$g_f(y) = \sup_{z \in \mathbb{R}^d} \langle z, y \rangle - f(z) = f^*(y), \quad \text{(convex conjugate)}$$

Semi-dual

$$\min_f \mathcal{S}(f) = \int f(x) \, \mathrm{d}P(x) + \int f^*(y) \, \mathrm{d}Q(y)$$

Non-linear, no constraints!

$$\int f^{**}(x) \, \mathrm{d}P(x) + \int f^*(y) \, \mathrm{d}Q(y), \quad T_0 = \nabla f_0$$

Upper bounds: empirical risk minimization

- Replace $\mathbb{E} \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n}$,

$$\mathcal{S}(f) = \int f(x)\, dP(x) + \int f^*(y)\, dQ(y) \rightsquigarrow \hat{\mathcal{S}}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) + \frac{1}{n}\sum_{i=1}^{n} f^*(Y_i)$$

- With hypothesis space $\mathcal{H} = V_J$ wavelet cutoff

$$f_0 = \arg\min_{f \in \mathcal{H}} \mathcal{S}(f), \quad \hat{f} = \arg\min_{f \in \mathcal{H}} \hat{\mathcal{S}}(f), \quad \hat{\mathrm{T}} = \hat{\nabla}\hat{f}$$

$$d(\hat{T}, T_0) \lesssim \mathcal{S}(\hat{f}) - \mathcal{S}(f_0) \lesssim \sup_{f \in \mathcal{H}} \left|\mathcal{S}(f) - \hat{\mathcal{S}}(f)\right| \quad \text{(generalization error)}$$

- To obtain upper bounds:
  1. Bound generalization error
  2. Relate excess risk to $d(\hat{T}, T_0)$: stability/margin condition

Semi-dual stability

Proposition (H., Rigollet)

Assume f, f^*, f_0 differentiable, f convex and $\nabla^2 f \preccurlyeq MI$ (M-Lipschitz). Then,

$$\int \|\nabla f(x) - \nabla f_0(x)\|_2^2 \, dP(x) \lesssim M \left(\mathcal{S}(f) - \mathcal{S}(f_0)\right),$$

$$\text{where} \quad \mathcal{S}(f) = \int f(x) \, dP(x) + \int f^*(y) \, dQ(y).$$

- Dualization of argument by Ambrosio (Gigli, 2011)
- If f additionally M-strongly convex, change roles to obtain

$$\int \|\nabla f(x) - \nabla f_0(x)\|_2^2 \, dP(x) \asymp \mathcal{S}(f) - \mathcal{S}(f_0)$$

More stability results

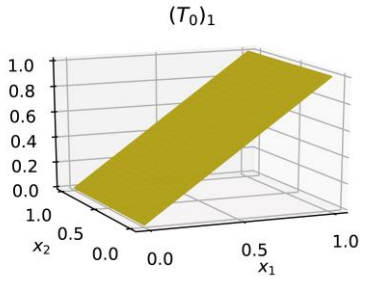
- Manole et al. 2021: Rates for smooth densities
  If
  - g_0 M-strongly convex and ∇g_0 M-Lipschitz
  - $\tilde{Q}$ any measure
  - $\tilde{f}$ Kantorovich potential for $(P, \tilde{Q})$

  Then,

$$\int \left\|\nabla \tilde{f}(x) - \nabla f_0(x)\right\|_2^2 dP(x) \lesssim W_2^2(P, \tilde{Q}) - W_2^2(P, Q) - \int g_0(y)\, d(\tilde{Q} - Q)(y)$$

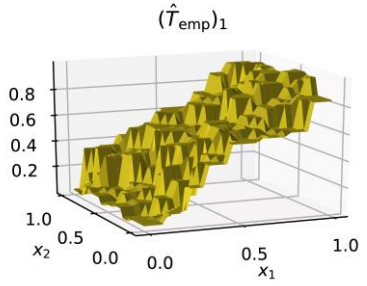
- Used to show plug-in rates for transport map estimation
- Two-sample case less general, needs regularity of candidate potential

Numerical results: estimators



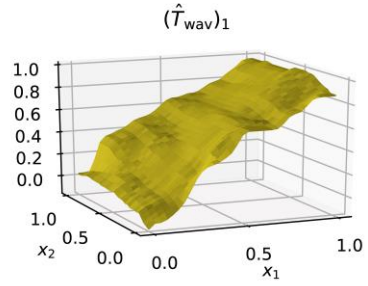
Ground truth

$T_0(x) = x$, identity map



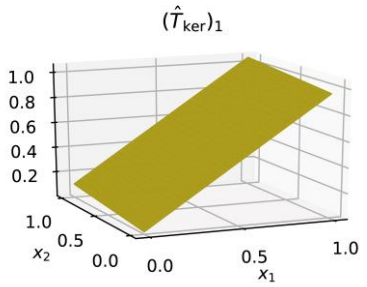
Baseline estimator $\hat{T}_{emp}$

$T_0(x) = x$, identity map



Wavelet estimator $\hat{T}_{wav}$

Discretize hyperbox around distribution, compute f^* with Linear-Time Legendre Transform, then interpolate. This is slow.



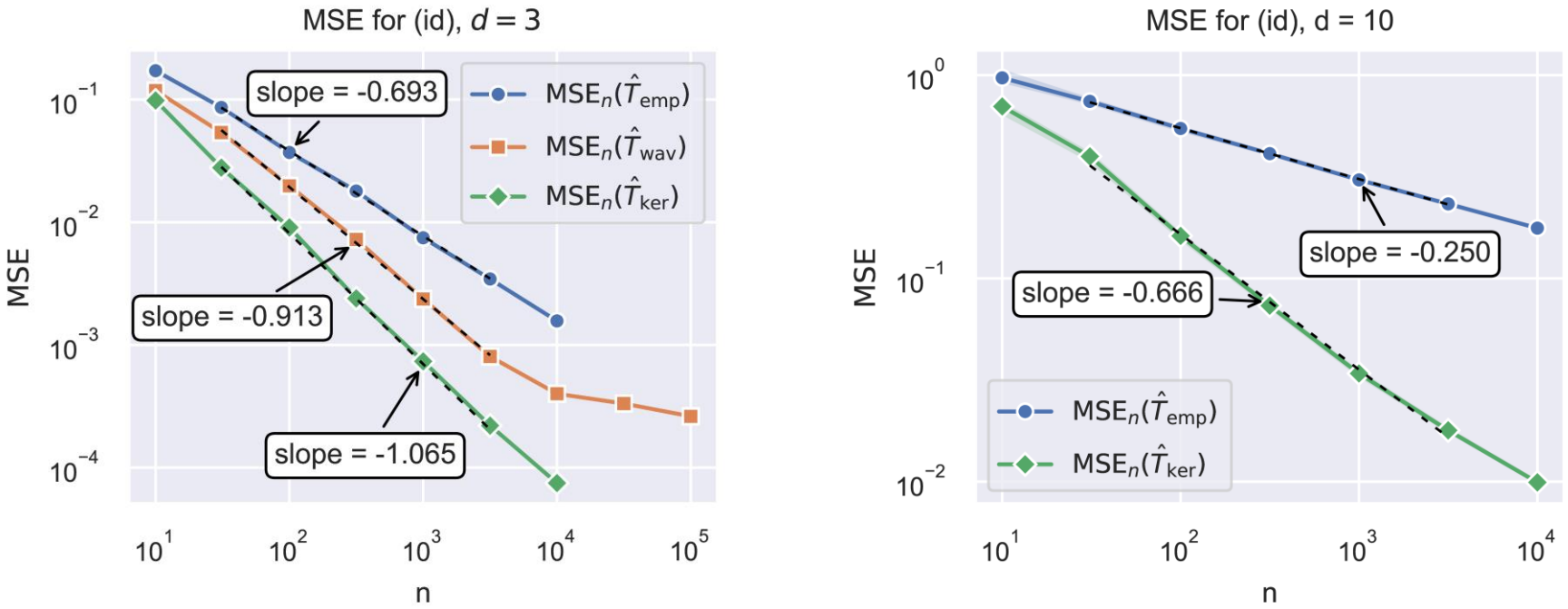
Heuristic kernel estimator $\hat{T}_{ker}$

Solve OT between $\hat{P}$, $\hat{Q}$ to get matching $\pi : [N] \to [N]$, compute kernel ridge estimator with input data $\left(X_i, Y_{\pi(i)}\right)_i$

Numerical results

Measure performance by

$$\mathrm{MSE}_n(\hat{T}) = \frac{1}{n}\sum_{i=1}^{n} \left\|\hat{T}(X_i) - T_0(X_i)\right\|_2^2$$



Smooth optimal transport maps

- Can mimic rates from classical non-parametric estimation for OT
- Smoothness can beat curse of dimensionality
- Stability estimates were useful
- Tricky: Get computationally efficient estimators

Entropic regularization provides smoothness, too

- Entropic regularization well-known for providing fast algorithms for OT
- Genevay et al. 2019, Mena, Weed, 2019:
  For compactly supported measures, can show that optimal potentials $f_\epsilon, g_\epsilon \in H^\alpha$ *with*
  $$\|f_\epsilon\|_{H^\alpha}, \|g_\epsilon\|_{H^\alpha} \lesssim 1 + \epsilon^{-1/(\alpha-1)}$$
- Obtain plug-in rates:
  $$\mathbb{E}\left[\left|W_\epsilon^2(\hat{P},\hat{Q}) - W_\epsilon^2(P,Q)\right|\right] \lesssim \frac{1}{\sqrt{n}}\left(1 + \epsilon^{-d/2}\right)$$
- Approximation guarantees
  $$\left|W_\epsilon^2(P,Q) - W_2^2(P,Q)\right| \lesssim \epsilon \log(\epsilon^{-1})$$
- Chizat et al., 2020: Improved approximation guarantees of order $\epsilon^2\, I(P,Q)$
- Still geometrically bad estimation rates for W_2^2

Primal

$$W_\epsilon^2(P,Q) = \min_\gamma \int \|y - x\|_2^2 \, d\gamma(x,y) + \epsilon D_{KL}(\gamma \,\|\, P \otimes Q)$$
$$\text{s.t. } \gamma \in \Gamma(P,Q),$$
$$D_{KL}(\gamma \,\|\, P \otimes Q) = \int \log\left(\frac{\gamma(x,y)}{P(x)Q(y)}\right) d\gamma(x,y)$$

Dual

$$(f_\epsilon, g_\epsilon) \in \max_{f,g} \int f(x)\, dP(x) + \int g(y)\, dQ(y) + \epsilon$$
$$-\epsilon \int \exp\left(\frac{f(x) + g(y) - \|y-x\|_2^2}{\epsilon}\right) dP(x)\, dQ(y)$$

Optimality condition

$$f_\epsilon(x) = -\epsilon \log \int \exp\left(\frac{g(y) - \|y-x\|_2^2}{\epsilon}\right) dQ(y)$$

Entropic OT also provides statistically efficient transport maps

- (Pooladian, Niles-Weed, 09/27/2021)
- Estimate transport map as barycentric projection of entropic plan, can write in terms of dual potentials (f_ϵ, g_ϵ):

$$\hat{T}_\epsilon(X_i) = \frac{\int y \, d\hat{\gamma}_\epsilon(X_i, y)}{\int d\hat{\gamma}_\epsilon(X_i, y)}, \quad \text{can extend to: } \hat{T}_\epsilon(x) = \frac{\int y \exp\left(\frac{g_\epsilon(y) - \|x-y\|_2^2}{\epsilon}\right) d\hat{Q}(y)}{\int \exp\left(\frac{g_\epsilon(y) - \|x-y\|_2^2}{\epsilon}\right) d\hat{Q}(y)}$$

- Under assumptions (including $\frac{1}{M} I \preccurlyeq \nabla^2 \mathrm{f} \preccurlyeq MI$), for $1 < \alpha \leq 3$ and $d' = 2\left\lceil \frac{d}{2} \right\rceil$,

$$\mathbb{E} \int \left\| \hat{T}_\epsilon(x) - T_0(x) \right\|_2^2 dP(x) \lesssim n^{-\frac{\alpha+1}{2\alpha+2+2d'}} \log(n)$$

- Compare to information-theoretic limit $n^{-\frac{2\alpha}{2\alpha-2+d}}$
- Goes through the same approximation steps as before for estimation of W_2^2, so won't beat geometrically bad dependence as $d \to \infty$

Computationally efficient estimator via RKHS

- Reproducing kernel Hilbert space
  - Hilbert space $\mathcal{H}$ of functions on $\mathbb{R}^d$
  - kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $f(x) = \langle k(x, .), f \rangle_{\mathcal{H}}$ for $f \in \mathcal{H}$
  - e.g. Sobolev space H^α if $\alpha > d/2$
  - Advantages
    - interpolation problems finite dimensional (representer theorem)
    - statistical guarantees
- Vacher et al, 2021: Under additional assumptions, if P, Q have H^α-densities with $\alpha > 3d$, obtain estimator $\widehat{W}_2^2(\hat{P}, \hat{Q})$ in time $O(n^2)$ with

$$\mathbb{E}\left[\left|\widehat{W}_2^2(\hat{P}, \hat{Q}) - W_2^2(P, Q)\right|\right] \lesssim n^{-\frac{1}{2}} \log(n)$$

- Compare to $n^{-\frac{2\alpha}{2\alpha-2+d}} \approx n^{-\frac{6}{7}}$ information-theoretic rate

$$\max \int f(x)\, d\hat{P}(x) + \int g(y)\, d\hat{Q}(y)$$

$$\text{s.t. } f(x) + g(y) \le \|x - y\|_2^2, \quad x, y \in \mathbb{R}^d$$

$$\Downarrow$$

$$\min \left\langle f, \frac{1}{n}\sum_{i=1}^n k(X_i, .) \right\rangle_{\mathcal{H}} + \left\langle g, \frac{1}{n}\sum_{j=1}^n k(Y_j, .) \right\rangle_{\mathcal{H}}$$

$$\text{s.t. } f, g \in \mathcal{H},$$

$$\|\tilde{X}_i - \tilde{Y}_i\|_2^2 - f(\tilde{X}_i) - g(\tilde{Y}_j) =$$

$$\langle k(\tilde{X}_i, .)k(\tilde{Y}_j, .), A[k(\tilde{X}_i, .)k(\tilde{Y}_j, .)] \rangle, \quad i, j \in [\ell]$$

for positive operator A and additional sample $\tilde{X}_i, \tilde{Y}_j$

Procedures without guarantees

FactoredOT for clustered distributions

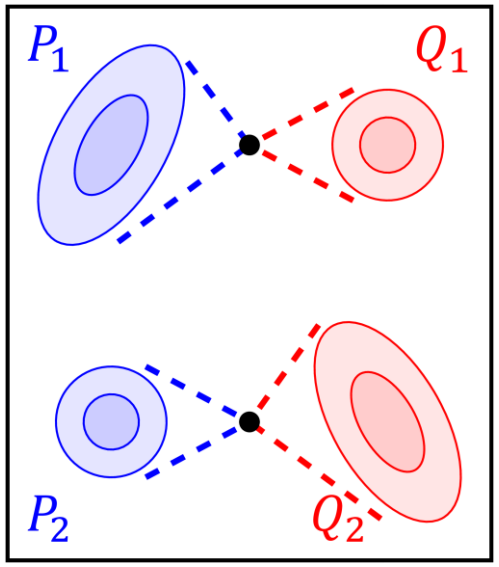
- Forrow, **H.**, et al., 2019:
  Gain statistical efficiency from sparse Wasserstein barycenters

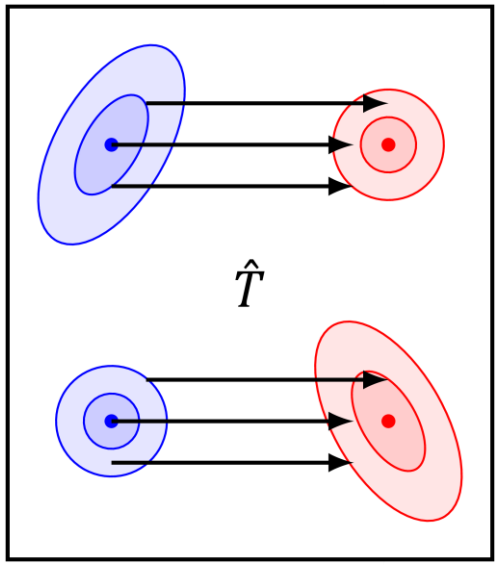
$$\hat{\rho}_k = \arg\min\{W_2^2(\hat{P}, \rho) + W_2^2(\rho, \hat{Q}) : \rho \text{ supported on } k \text{ points}\}$$

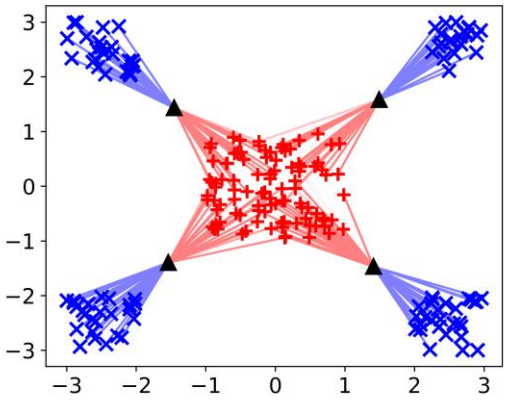
- Generalization bounds

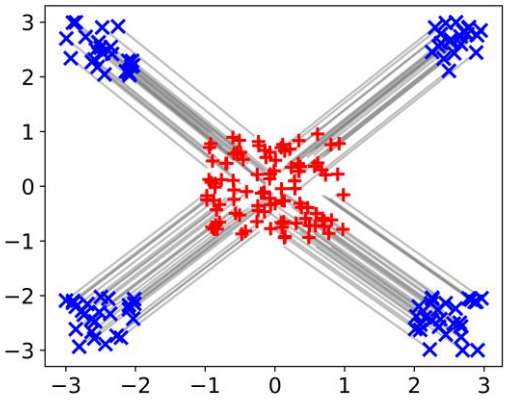
$$\sup_{\rho} \mathbb{E}\left[\left|W_2^2(\rho, P) - W_2^2(\rho, \hat{P})\right|\right] \lesssim \frac{\sqrt{k^3 \log k}}{\sqrt{n}}$$

- But no approximation guarantees









Nearest Brenier potential

- Paty, d'Aspremont, Cuturi, 2020:
  Alternative way of ensuring strong convexity:

$$\hat{f}_{SSNB} = \arg\min\{W_2^2(\nabla f_{\#}\hat{P}, \hat{Q}) : f \text{ strongly convex and } \nabla f \text{ Lipschitz}\}$$

- Leads to mixed quadratically constrained quadratic program/Wasserstein problem

Input convex neural networks

- Recall semi-dual:

$$\mathcal{S}(f) = \int f(x)\, dP(x) + \int f^*(y)\, dQ(y)$$

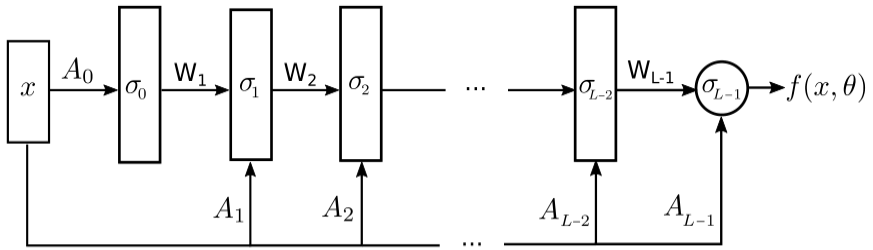
$$\begin{aligned} f^*(y) &= \sup_x \langle x, y\rangle - f(x) \\ &= \langle \nabla f^*(y), y\rangle - f\big(\nabla f^*(y)\big) \\ &\geq \langle y, \nabla g(y)\rangle - f\big(\nabla g(y)\big) \quad \forall g \end{aligned}$$

- Makkuva et al., 2020: Replace f, g by input-convex neural networks, $ICNN$:

$$\min_{f \in ICNN} \max_{g \in ICNN} \int f(x)\, d\hat{P}(x) + \int \big[\langle y, \nabla g(y)\rangle - f\big(\nabla \mathrm{g}(\mathrm{y})\big)\big]\, d\hat{Q}(y)$$

Input convex neural network:

$$\begin{aligned} z_{l+1} &= \sigma_l(W_l z_l + A_l x + b_l), \\ W_l &\geq 0, \quad \text{for } l = 0, \ldots, L-1 \\ f(x) &= z_L \end{aligned}$$



Conclusion

- OT can be very useful, but also very noisy
- Compromise: projection robust Wasserstein distance, entropic OT
- Regularize: recent progress in regularized estimators with statistical and computational guarantees
- Improve:
  - Computationally efficient estimator matching information-theoretic rates
  - Rates for plug-in transport map estimators with smoothing

Thanks!

References

- Chizat, Lenaic, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. “Faster Wasserstein Distance Estimation with the Sinkhorn Divergence.” *Advances in Neural Information Processing Systems* 33 (2020).
- Courty, Nicolas, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. “Optimal Transport for Domain Adaptation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, no. 9 (2016): 1853–65.
- Deb, Nabarun, Promit Ghosal, and Bodhisattva Sen. “Rates of Estimation of Optimal Transport Maps Using Plug-in Estimators via Barycentric Projections.” *ArXiv:2107.01718 [Math, Stat]*, July 4, 2021. http://arxiv.org/abs/2107.01718.
- Dudley, Richard Mansfield. “The Speed of Mean Glivenko-Cantelli Convergence.” *The Annals of Mathematical Statistics* 40, no. 1 (1969): 40–50.

References

- Arjovsky, Martin, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks.” In *International Conference on Machine Learning*, 214–23. PMLR, 2017.
- Forrow, Aden, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. “Statistical Optimal Transport via Factored Couplings.” In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2454–65. PMLR, 2019. https://proceedings.mlr.press/v89/forrow19a.html.
- Fournier, Nicolas, and Arnaud Guillin. “On the Rate of Convergence in Wasserstein Distance of the Empirical Measure.” *Probability Theory and Related Fields* 162, no. 3 (2015): 707–38.
- Frogner, Charlie, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. “Learning with a Wasserstein Loss.” *ArXiv Preprint ArXiv:1506.05439*, 2015.
- Genevay, Aude, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. “Sample Complexity of Sinkhorn Divergences.” In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1574–83. PMLR, 2019. https://proceedings.mlr.press/v89/genevay19a.html.
- Hütter, Jan-Christian, and Philippe Rigollet. “Minimax Estimation of Smooth Optimal Transport Maps.” *The Annals of Statistics* 49, no. 2 (April 2021): 1166–94. https://doi.org/10.1214/20-AOS1997.

References

- Kolouri, Soheil, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. “Generalized Sliced Wasserstein Distances.” In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper/2019/hash/f0935e4cd5920aa6c7c996a5ee53a70f-Abstract.html.
- Li, Wenbo, and Ricardo H. Nochetto. “Quantitative Stability and Error Estimates for Optimal Transport Plans.” *IMA Journal of Numerical Analysis* 41, no. 3 (2021): 1941–65.
- Lin, Tianyi, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael I. Jordan. “Projection Robust Wasserstein Distance and Riemannian Optimization.” *ArXiv Preprint ArXiv:2006.07458*, 2020.
- Lin, Tianyi, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael Jordan. “On Projection Robust Optimal Transport: Sample Complexity and Model Misspecification.” In *International Conference on Artificial Intelligence and Statistics*, 262–70. PMLR, 2021.
- Makkuva, Ashok, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. “Optimal Transport Mapping via Input Convex Neural Networks.” In *Proceedings of the 37th International Conference on Machine Learning*, 6672–81. PMLR, 2020. https://proceedings.mlr.press/v119/makkuva20a.html.

References

- Manole, Tudor, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. “Plugin Estimation of Smooth Optimal Transport Maps.” *ArXiv Preprint ArXiv:2107.12364*, 2021.
- Manole, Tudor, and Jonathan Niles-Weed. “Sharp Convergence Rates for Empirical Optimal Transport with Smooth Costs.” *ArXiv Preprint ArXiv:2106.13181*, 2021.
- Mena, Gonzalo, and Jonathan Weed. “Statistical Bounds for Entropic Optimal Transport: Sample Complexity and the Central Limit Theorem.” *ArXiv:1905.11882 [Cs, Math, Stat]*, May 30, 2019. http://arxiv.org/abs/1905.11882.
- Niles-Weed, Jonathan, and Philippe Rigollet. “Estimation of Wasserstein Distances in the Spiked Transport Model.” *ArXiv Preprint ArXiv:1909.07513*, 2019.
- Paty, François-Pierre, Alexandre d’Aspremont, and Marco Cuturi. “Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport.” In *International Conference on Artificial Intelligence and Statistics*, 1222–32. PMLR, 2020.

References

- Pooladian, Aram-Alexandre, and Jonathan Niles-Weed. “Entropic Estimation of Optimal Transport Maps.” *ArXiv:2109.12004 [Math, Stat]*, September 24, 2021. http://arxiv.org/abs/2109.12004.
- Rabin, Julien, Julie Delon, and Yann Gousseau. “Regularization of Transportation Maps for Color and Contrast Transfer.” In *2010 IEEE International Conference on Image Processing*, 1933–36. IEEE, 2010.
- Rolet, Antoine, Marco Cuturi, and Gabriel Peyré. “Fast Dictionary Learning with a Smoothed Wasserstein Loss.” In *Artificial Intelligence and Statistics*, 630–38. PMLR, 2016.
- Schiebinger, Geoffrey, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Siyan Liu, Stacie Lin, Peter Berube, and Lia Lee. “Reconstruction of Developmental Landscapes by Optimal-Transport Analysis of Single-Cell Gene Expression Sheds Light on Cellular Reprogramming.” *BioRxiv*, 2017, 191056.
- Vacher, Adrien, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. “A Dimension-Free Computational Upper-Bound for Smooth Optimal Transport Estimation.” *ArXiv Preprint ArXiv:2101.05380*, 2021.
- Weed, Jonathan, and Francis Bach. “Sharp Asymptotic and Finite-Sample Rates of Convergence of Empirical Measures in Wasserstein Distance.” *Bernoulli* 25, no. 4A (2019): 2620–48.
- Weed, Jonathan, and Quentin Berthet. “Estimation of Smooth Densities in Wasserstein Distance.” In *Conference on Learning Theory*, 3118–19. PMLR, 2019.

Appendix

Proof of semi-dual stability

convex conjugate: $f^*(y) = \sup_{z \in \mathbb{R}^d} \langle y, z \rangle - f(z)$

$f^*(y) = \langle y, x \rangle - f(x) \Leftrightarrow y = \nabla f(x)$

f convex, ∇f $L-$Lipschitz: $f^*(y) \geq f^*(x) + \langle \nabla f^*(x), y - x \rangle + \frac{1}{2L} \|y - x\|_2^2$

$\nabla f^* = (\nabla f)^{-1}$

$$\int f(x)\,\mathrm{d}P(x) + \int f^*(y)\,\mathrm{d}Q(y)$$

$$= \int \left[f(x) + f^*\left(\nabla f_0(x)\right) \right] \mathrm{d}P(x), \quad (Q = (T_0)_{\#}P = (\nabla f_0)_{\#}P)$$

$$\geq \int \left[f(x) + \underbrace{f^*\left(\nabla f(x)\right)}_{= \langle x, \nabla f(x) \rangle - f(x)} + \left\langle \underbrace{\nabla f^*\left(\nabla f(x)\right)}_{x}, \nabla f_0(x) - \nabla f(x) \right\rangle + \frac{1}{2L} \|\nabla f_0(x) - \nabla f(x)\|_2^2 \,\mathrm{d}P(x) \right]$$

$$= \int \left[\langle x, \nabla f_0(x) \rangle + \frac{1}{2L} \|\nabla f_0(x) - \nabla f(x)\|_2^2 \,\mathrm{d}P(x) \right]$$

$$= \int f_0(x)\,\mathrm{d}P(x) + \int f_0^*(y)\,\mathrm{d}Q(y) + \frac{1}{2L} \int \|\nabla f_0(x) - \nabla f(x)\|_2^2 \,\mathrm{d}P(x)$$