Statistical Inference for Linear Programming Problems: Debiasing and Regularity

Arun Kumar Kuchibhotla

Carnegie Mellon University

3 November, 2026

Collaborators

Koustav Goswami (Indian Statistical Institute)
Hannes Leeb (University of Vienna)

Table of contents

1. Discrete OT: An Example and Existing Results
2. Let's Zoom: Pitfalls of Smoothing
3. Two Impossibilities

Outline

Discrete OT: An Example and Existing Results

Let's Zoom: Pitfalls of Smoothing

Two Impossibilities

Optimal Transport

⋆ Optimal transport has a rich history of over 150 years in mathematics and operations research. It stemmed from the problem of moving dirt from one place to another with minimal cost of transport.

⋆ Given two distributions μ and ν on $\mathbb{R}^d$, a map $T : \mathbb{R}^d \to \mathbb{R}^d$ is said to be a transport map if $T(X) \sim \nu$ when $X \sim \mu$. Optimal transport is defined as the transport map that minimizes a certain cost. For example, with squared Euclidean loss, the optimal transport map is

$$T^* := \underset{T:T(X)\sim\nu, X\sim\mu}{\arg\min} \mathbb{E}\|T(X) - X\|^2.$$

⋆ A deterministic transport map may not always exist, and moreover, the non-convexity of the constraint make this problem difficult to solve. An alternative is to consider couplings:

$$\pi^* := \arg\min_{\pi} \int \|x - y\|^2 d\pi(x, y),$$

over all joint distributions $\pi(\cdot, \cdot)$ such that $\pi(A, \mathbb{R}^d) = \mu(A)$ and $\pi(\mathbb{R}^d, B) = \nu(B)$.

Discrete OT

⋆ Suppose μ and ν are two distributions supported on $\{1, 2, \ldots, m\}$ for some $m < \infty$. Let

$$\mu(i) = t_i \quad \text{and} \quad \nu(i) = s_i,$$

so that $t_i, s_i \in [0, 1]$ and $\sum_{i=1}^m t_i = 1$ and $\sum_{i=1}^m s_i = 1$.

⋆ The discrete OT problem now becomes

$$\pi^* = \underset{\pi \in [0,1]^{m \times m}}{\arg\min} \sum_{1 \le i,j \le m} c_{ij}\pi_{ij} \quad \text{subject to} \quad \pi_{ij} \ge 0, \sum_{i=1}^m \pi_{ij} = s_j, \sum_{j=1}^m \pi_{ij} = t_i.$$

Here c_{ij} is the cost of transport i to j.

⋆ Given only data from μ and ν, the sample version of discrete OT is

$$\widehat{\pi}_n = \underset{\pi \in [0,1]^{m \times m}}{\arg\min} \sum_{1 \le i,j \le m} c_{ij}\pi_{ij} \quad \text{subject to} \quad \pi_{ij} \ge 0, \sum_{i=1}^m \pi_{ij} = \widehat{s}_j, \sum_{j=1}^m \pi_{ij} = \widehat{t}_i.$$

An Example

⋆ Suppose $m = 2$ (i.e., distributions supported on two points $\{1, 2\}$). Let $c_{ij} = \mathbf{1}\{i \neq j\}$.

⋆ Then

$$\pi^* = \arg\min_{\pi} \ \pi_{12} + \pi_{21} \quad \text{subject to} \quad \pi \geq 0, \sum_{i=1}^{2} \pi_{ij} = s_j, \sum_{j=1}^{2} \pi_{ij} = t_i.$$

⋆ If $\mathbf{t} = \mathbf{s} = (1/2, 1/2)$ (self-transport) and $\widehat{\mathbf{t}}, \widehat{\mathbf{s}}$ are empirical probability vectors, then

$$n^{1/2}(\widehat{\pi}_n - \pi^*) \ \xrightarrow{d} \ \frac{1}{2} \begin{bmatrix} \min\{Z_1, Z_2\} & (Z_1 - Z_2)_+ \\ (Z_2 - Z_1)_+ & -\max\{Z_1, Z_2\} \end{bmatrix},$$

where Z_1, Z_2 are independent standard Gaussian random variables.

⋆ Such non-normal limiting distributions are common under constraints.

⋆ Small perturbations of the linear constraints can lead to abrupt changes in the geometry of the constraint set, which in turn gives rise to non-differentiable changes in the optimal solution.

A General Setting

⋆ Note that discrete OT is a specific instance of linear programs. One can consider a more general problem:

$$\theta^* := \arg\min_{\theta} c^\top \theta \quad \text{subject to} \quad A\theta = b^*, \theta \geq 0.$$

⋆ In discrete OT, c, A are known for both the population and sample problems, and b^* is estimated from the sample. Hence, the sample version is

$$\widehat{\theta}_n := \arg\min_{\theta} c^\top \theta \quad \text{subject to} \quad A\theta = \widehat{b}_n, \theta \geq 0.$$

⋆ More generally, the collection of problems can be summarized as

$$\theta^*(b) := \arg\min_{\theta} c^\top \theta \quad \text{subject to} \quad A\theta = b, \theta \geq 0.$$

⋆ It is important to note that if $\theta^*(b^*)$ is uniquely defined for some b^*, then it is not necessary that $\theta^*(b)$ is uniquely defined for b close to b^*. This already hints at some irregular behavior.

Non-differentiability and Smoothing

- ⋆ If the map $b \mapsto \theta^*(b)$ is differentiable at b^*, then by the delta method, we get

$$n^{1/2}(\widehat{\theta}_n - \theta^*) \xrightarrow{d} N(0, \Sigma) \quad \text{whenever} \quad n^{1/2}(\widehat{b}_n - b^*) \xrightarrow{d} N(0, \Gamma),$$

  for some covariance matrices Σ, Γ.

- ⋆ The discrete OT example with $\mathbf{t} = \mathbf{s} = (1/2, 1/2)$ shows that differentiability may not hold.

- ⋆ In fact, $b \mapsto \theta^*(b)$ is piecewise linear, and hence, Lipschitz continuous, but may be non-differentiable.

- ⋆ Non-normality of the limiting distributions makes it difficult to perform statistical inference, i.e., construction of confidence sets for θ^* based on the data.

- ⋆ A common method in stochastic programming (e.g., Juditsky and Nemirovski, 2020) is to assume an upper bound on the Lipschitz constant and convert confidence sets for b^* to confidence sets for θ^*. These can be severely conservative.

- ⋆ Another common method is to smooth the map $b \mapsto \theta^*(b)$.

Smoothing

⋆ Recall

$$\theta^*(b) := \arg\min_{\theta} \; c^\top \theta \quad \text{subject to} \quad A\theta = b, \theta \geq 0$$

$$= \arg\min_{\theta} \; c^\top \theta + \mathbb{I}\{A\theta = b, \theta \geq 0\}, \quad \text{where} \quad \mathbb{I}(C) = \begin{cases} 0, & \text{if } \theta \in C, \\ \infty, & \text{otherwise.} \end{cases}$$

⋆ Non-differentiability mainly arises from non-smoothness of $\mathbb{I}(\cdot)$. For a Gaussian limiting distribution, researchers have suggested considering a smoothed target.

⋆ Consider

$$\theta^*(b; \lambda) := \arg\min_{\theta:\, A\theta = b} \; c^\top \theta + p_\lambda(\theta),$$

where $\theta \mapsto p_\lambda(\theta)$ for $\lambda > 0$ is a smooth differentiable real-valued approximation of $\theta \mapsto \mathbb{I}(\theta \geq 0)$. For example, one could use the log-barrier function.

Smoothing and Debiasing

⋆ The readily apparent issue with this is that

$$\theta^*(b;\lambda) \neq \theta^*(b) \quad \text{for} \quad \lambda > 0.$$

⋆ One can view $\lambda \mapsto \theta^*(b;\lambda)$ itself as a perturbation problem and study its behavior.

⋆ Under certain smoothness assumptions on $(\lambda, \theta) \mapsto p_\lambda(\theta)$, it can be proved that $\lambda \to \theta^*(b, \lambda)$ is differentiable at $\lambda = 0$. This has been used to create debiased targets.

⋆ Formally,

$$\begin{aligned} \theta^*(b;\lambda) &= \theta^*(b) + \lambda d + O(\lambda^2) \quad \text{as} \quad \lambda \to 0, \\ \Rightarrow \quad 2\theta^*(b;\lambda/2) - \theta^*(b;\lambda) &= \theta^*(b) + O(\lambda^2) \quad \text{as} \quad \lambda \to 0. \end{aligned}$$

⋆ Hence for a proper choice of $\lambda = \lambda_n$, we get $\widehat{\theta}_n = 2\theta^*(\widehat{b};\lambda/2) - \theta^*(\widehat{b};\lambda)$ satisfies

$$n^{1/2}(\widehat{\theta}_n - \theta^*(b^*)) \xrightarrow{d} N(0, \Sigma),$$

for some covariance matrix Σ.

Smoothing and Debiasing

- ⋆ Liu et al. (2025, arXiv:2505.04312) applied this Richardson extrapolation idea and proved asymptotic normality of the estimator.
- ⋆ They also proposed confidence sets for $\theta^*(b^*)$ using bootstrap of $\widehat{\theta}_n$.
- ⋆ For simplicity, assume $n^{1/2}(\widehat{b}_n - b^*) \sim N(0, I_d)$, then their result implies that

$$\lim_{n\to\infty} |\mathbb{P}_{b^*}(\theta^*(b^*) \in \widehat{\mathrm{CI}}_{n,\alpha}) - (1-\alpha)| = 0 \quad \text{for all} \quad b^*.$$

- ⋆ For the example of discrete OT with $\mathbf{t} = \mathbf{s} = (1/2, 1/2)$, we get that

$$n^{1/2}(\widehat{\pi}_{ij} - \pi_{ij}) \xrightarrow{d} N(0, \sigma_{ij}^2) \quad \text{as} \quad n \to \infty,$$

and their confidence intervals remain valid as $n \to \infty$.

Outline

Discrete OT: An Example and Existing Results

Let's Zoom: Pitfalls of Smoothing

Two Impossibilities

Instead of $\mathbf{t} = \mathbf{s} = (1/2, 1/2)$, let us consider two different but close enough distributions:

$$\mathbf{t} = \begin{bmatrix} 1/2 + n^{-1/2}u & 1/2 - n^{-1/2}u \end{bmatrix} \quad \text{and} \quad \mathbf{s} = \begin{bmatrix} 1/2 + n^{-1/2}v & 1/2 - n^{-1/2}v \end{bmatrix}.$$

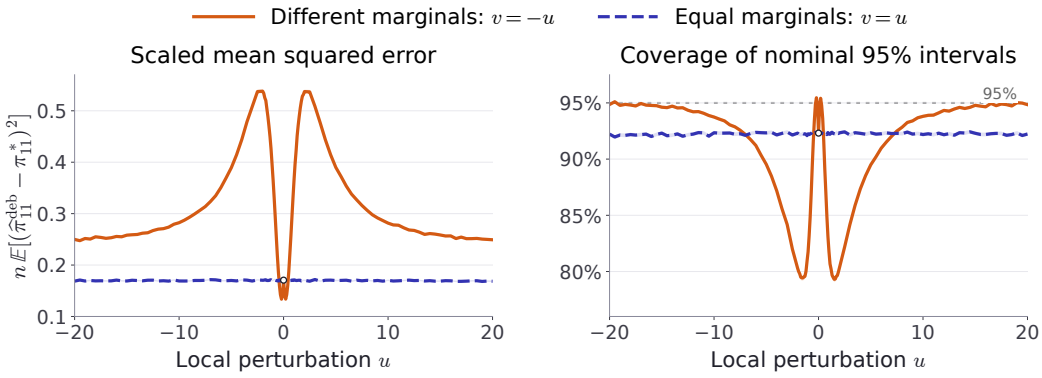


Figure: Scaled MSE and coverage of nominal 95% Wald intervals for the Richardson-debiased log-barrier estimator of π_{11}^*. $n = 10^5$, $\lambda = n^{-1/3}$; 10^5 replications.

Pitfalls of Smoothing

⋆ In fact, from the directional differentiability of $b \mapsto \theta^*(b)$, it follows that

$$n^{1/2}(\widehat{\theta}_n - \theta^*(b^* + h/\sqrt{n})) \xrightarrow{d} N(Mh - \nabla\theta^*(b^*)[h], \Sigma),$$

where $\nabla\theta^*(b^*)[h]$ is the directional derivative of $b \mapsto \theta^*(b)$ in the direction of h. Hence, unless the directional derivative is linear at b^*, there is bias in the limiting distribution.

⋆ This implies that for

$$\mathbf{t} = \begin{bmatrix} 1/2 + n^{-1/2}u & 1/2 - n^{-1/2}u \end{bmatrix} \quad \text{and} \quad \mathbf{s} = \begin{bmatrix} 1/2 + n^{-1/2}v & 1/2 - n^{-1/2}v \end{bmatrix},$$

we get

$$\lim_{n\to\infty} \sup_{u+v=0, |u|\leq 1} |\mathbb{P}_{\mathbf{t},\mathbf{s}}(\pi^* \in \widehat{\mathrm{CI}}_{n,\alpha}) - (1-\alpha)| \neq 0.$$

It is pointwise valid but not locally uniformly.

⋆ Note that one can obtain uniformly valid inference using Lipschitzness of $b \mapsto \theta^*(b)$.

Outline

Discrete OT: An Example and Existing Results

Let's Zoom: Pitfalls of Smoothing

Two Impossibilities

Summary and General Result

- ⋆ We have shown that the Richardson extrapolation idea does not yield uniformly valid confidence sets for linear programs, and in particular, for discrete OT.
- ⋆ One may ask why should we care about uniform validity. Well ...
- ⋆ Similar issues also arise when estimating optimal value of a stochastic program. It is well-known (Shapiro, 2021) that

$$n^{1/2}\left(\inf_{\theta\in\Theta}\widehat{P}_n[\ell(X,\theta)] - \inf_{\theta\in\Theta}P[\ell(X,\theta)]\right) \xrightarrow{d} \inf_{\theta\in\Theta_{\min}} G_\theta,$$

  for a Gaussian process G_θ and $\Theta_{\min}$ is the set of minimizers in the population.

- ⋆ This is a pointwise result and when the distribution shifts slightly to yield a significant change in $\Theta_{\min}$, then pointwise results can be misleading.

General Result

⋆ We have the following general result for inference about $\theta^*(b)$.

⋆ Let $\mathcal{B} \subseteq \mathbb{R}^m$ be an open set. Suppose we observe $\widehat{b}_n \sim N(b, I_m/n)$ for some $b \in \mathcal{B}$ and we are interested in inference for

$$\theta^*(b) := \underset{\theta \in \mathbb{R}^d}{\arg\min}\ c^\top \theta \quad \text{subject to} \quad A\theta = b, \theta \geq 0.$$

⋆ Note that the plug-in estimator $\widehat{\theta}_n = \theta^*(\widehat{b}_n)$ satisfies

$$\widehat{\theta}_n - \theta^*(b) = O_p(n^{-1/2}), \quad \text{i.e., minimax rate optimal,}$$

uniformly over $b \in \mathcal{B}$. Lipschitzness also implies Juditsky and Nemirovski approach of inference is also uniformly valid (but is conservative).

General Result 1

⋆ If $b^* \in \mathcal{B}$ is such that $b \mapsto \theta^*(b)$ is not differentiable at b^*, then there *cannot* exist an estimator T_n such that

$$n^{1/2}\big(T_n - \theta^*(b^*)\big) = O_p(1), \quad \text{as} \quad n \to \infty,$$

and an (estimated) probability measure $\widehat{\mu}_n$ such that

$$d_{\text{Levy}}(\widehat{\mu}_n,\, \mu_{b^*+h/\sqrt{n}}) \xrightarrow[P_{b^*+h/\sqrt{n}}]{p} 0 \quad \text{as} \quad n \to \infty \quad \text{for all} \quad \|h\| \leq 1.$$

Here $\mu_{b^*+h/\sqrt{n}}$ is the probability measure of $n^{1/2}(T_n - \theta^*(b^* + h/\sqrt{n}))$ under $b^* + h/\sqrt{n}$.

⋆ This implies that no resampling method can consistently estimate the finite-sample distribution of a minimax rate-optimal estimator.

⋆ In particular, for any estimator either it is rate suboptimal or it is optimal and no bootstrap/subsampling method can be uniformly valid.

General Result 2

⋆ If $b^* \in \mathcal{B}$ is such that $b \mapsto \theta^*(b)$ is not differentiable at b^*, then for several directions u, there cannot exist an estimator T_n such that

$$n^{1/2}(T_n - u^\top \theta^*(b^*)) = O_p(1), \quad \text{as} \quad n \to \infty,$$

and is locally asymptotically median unbiased, i.e.,

$$\liminf_{n\to\infty} \min \left\{ \begin{array}{l} \mathbb{P}_{b^*+h/\sqrt{n}} \left(T_n \geq u^\top \theta^*(b^* + h/\sqrt{n}) \right), \\ \mathbb{P}_{b^*+h/\sqrt{n}} \left(T_n \leq u^\top \theta^*(b^* + h/\sqrt{n}) \right) \end{array} \right\} \geq \frac{1}{2} \quad \text{for all} \quad \|h\| \leq 1.$$

⋆ Consequently, there cannot exist (equal-tailed or one-sided) confidence intervals for $u^\top \theta^*(\cdot)$ that are locally asymptotically *exact* and simultaneously minimax optimal.

⋆ Hence, no smoothing or other approximations and no bootstrap or modifications can yield "good" estimators or "good" confidence intervals for $\theta^*(\cdot)$.

Conclusions

- ⋆ Stochastic programs with constraints or non-smooth components are ubiquitous in Statistics, Econometrics, and Operations research.
- ⋆ Traditional techniques for inference in these problems rely on smoothing and debiasing in order to get asymptotically normal estimator.
- ⋆ Non-smooth stochastic programs trigger non-differentiability of the minimizers in perturbed problems and we showed that this makes inference difficult.
- ⋆ For linear programs specifically, the Lipschitz constant can be obtained using Hoffman's lemma that does not depend on b. Hence, conservative inference is possible and such conservativeness cannot be avoided.
- ⋆ In general, non-differentiability in perturbed problems implies no bootstrap or resampling technique can yield (asymptotically) uniform consistency in distribution estimation.