

# Solving a Research Problem in Mathematical Statistics with AI Assistance

Edgar Dobriban\*

November 25, 2025

## Abstract

Over the last few months, AI models including large language models have improved greatly. There are now several documented examples where they have helped professional mathematical scientists prove new results, sometimes even helping resolve known open problems. In this short note, we add another example to the list, by documenting how we were able to solve a previously unsolved research problem in robust mathematical statistics with crucial help from GPT-5. Our problem concerns robust density estimation, where the observations are perturbed by Wasserstein-bounded contaminations. In a previous preprint (Chao and Dobriban, 2023, arxiv:2308.01853v2), we have obtained upper and lower bounds on the minimax optimal estimation error; which were, however, not sharp.

Starting in October 2025, making significant use of GPT-5 Pro, we were able to derive the minimax optimal error rate (reported in version 3 of the above arxiv preprint). GPT-5 provided crucial help along the way, including by suggesting calculations that we did not think of, and techniques that were not familiar to us, such as the dynamic Benamou-Brenier formulation, for key steps in the analysis. Working with GPT-5 took a few weeks of effort, and we estimate that it could have taken several months to get the same results otherwise. At the same time, there are still areas where working with GPT-5 was challenging: it sometimes provided incorrect references, and glossed over details that sometimes took days of work to fill in. We outline our workflow and steps taken to mitigate issues. Overall, our work can serve as additional documentation for a new age of human-AI collaborative work in mathematical science.

## 1 Introduction

Over the last few months, AI models for text generation have improved greatly, with new releases in major families, including GPT 5, Gemini, Claude, etc. Among a variety of use cases, there are now several documented examples where they have helped professional mathematical scientists prove new results, sometimes even helping resolve known open problems, see e.g., Feldman and Karbasi (2025); Jang and Ryu (2025); Salim (2025); Bubeck et al. (2025), etc.

In this short note, we add another example to the list, by documenting how we were able to solve a previously unsolved research problem in robust mathematical statistics—where publicly documented examples of AI-assisted proofs are relatively more sparse—with crucial help from GPT-5. Our problem concerns robust density estimation when the observations are perturbed

---

\*Department of Statistics and Data Science, University of Pennsylvania. E-mail address:  
[dobriban@wharton.upenn.edu](mailto:dobriban@wharton.upenn.edu).

by Wasserstein-bounded contaminations. In a previous preprint (Chao and Dobriban, 2023, v2), we have obtained upper and lower bounds on the minimax optimal estimation error; which were, however, not sharp.

Starting in October 2025, making significant use of GPT-5 Pro, we were able to derive the minimax optimal error rate, showing that neither of the previous upper/lower bounds were tight. GPT-5 provided crucial help along the way, including by suggesting techniques that were not familiar to us, such as the dynamic Benamou-Brenier formulation (Brenier, 2004), for key steps in the analysis.

Working with GPT-5 took a few weeks of effort, and we estimate that it could have taken several months to get the same results otherwise. At the same time, there are still areas where working with GPT-5 was challenging: it provided incorrect references, and glossed over details that sometimes took days of work to fill in. We outline our workflow and steps taken to mitigate issues. Overall, our work can serve as additional documentation for a new age of human-AI collaborative work in mathematical science.

## 2 The Problem: Robust Density Estimation under Wasserstein Contaminations

In this section, we aim to describe the problem that we have considered at a level that should be accessible to those familiar with undergraduate multivariable calculus, probability theory, and statistics.

Our problem<sup>1</sup> is the following:

We observe an i.i.d. sample  $X'_1, \dots, X'_n$  in  $\mathbb{R}^p$  from a density  $\tilde{f}$ . However, this sample has been contaminated. This density  $\tilde{f}$  of the data has been perturbed by a bounded contamination compared to the true density  $f$ . How well can we recover the true density based on the contaminated data?

**Wasserstein contaminations.** The level of perturbation is assumed to be captured by a Wasserstein norm. Specifically, consider some  $\ell_q$  norm,  $q \in [1, \infty]$ . A typical case is the Euclidean norm, where  $q = 2$ . Let  $X_1, \dots, X_n$  in  $\mathbb{R}^p$  be an i.i.d. sample from the true density  $f$ . Let  $\varepsilon \geq 0$  represent the strength of the perturbation. Then we assume that, for some  $r \geq 1$ , there is a joint realization of the original and contaminated data, such that  $(\mathbb{E}\|X'_i - X_i\|_q^r)^{1/r} \leq \varepsilon$ , for all  $i = 1, \dots, n$ . Denoting by  $W_{q,r}$  the Wasserstein- $r$  metric induced by the  $\ell_q$  norm on  $\mathbb{R}^p$ , this corresponds to<sup>2</sup>  $W_{q,r}(X'_i, X_i) \leq \varepsilon$ , see e.g., Villani (2003, 2008); Santambrogio (2015); Figalli and Glaudo (2021), etc.

Now consider an arbitrary given  $x_0 \in \mathbb{R}^p$ . How well can we estimate the true density  $f(x_0)$  based on the contaminated sample? And what is the best density estimator? To quantify the level of error, we can choose a loss function,<sup>3</sup> such as the squared error  $L(\hat{f}, f) = (\hat{f}(x_0) - f(x_0))^2$ .

---

<sup>1</sup>To the best of our knowledge, our question is indeed an unsolved problem. However, it is also not clear if anybody has formally proposed this precise problem before (beyond our own prior work on the problem). Thus, perhaps one might not call it an open problem. However, this reflects a dominant mode of thinking and research in statistics, where open problems have rarely been the main drivers of progress; see, for instance Wasserman (2012); Jordan (2011) for some informal discussion.

<sup>2</sup>The rigorous definition of the metric is on probability distributions instead of random variables. We use an informal definition here for simplicity.

<sup>3</sup>Our result actually holds for much more general loss functions of the form  $\Phi(|\hat{f}(x_0) - f(x_0)|)$ , where  $\Phi$  is a real-valued, non-negative, convex function that maps zero to zero and does not grow faster than an exponential, satisfying a doubling property, see Chao and Dobriban (2023). However, we will focus on the squared loss for clarity.

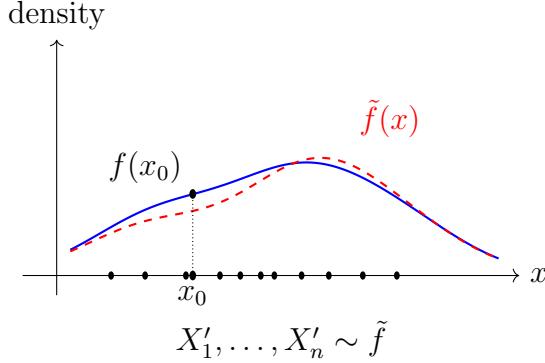


Figure 1: True density  $f$ , contaminated density  $\tilde{f}$ , observed contaminated sample  $X'_1, \dots, X'_n$ , and the target value  $f(x_0)$  to be estimated.

This contamination model corresponds to small but systematic perturbations of the entire original sample, and can be contrasted to the Huber  $\varepsilon$ -contamination model, which allows large, potentially unbounded perturbations of a small fraction of the data, see e.g., Huber (1964, 1965, 2004); Hampel et al. (2005), etc. The study of the Wasserstein contamination model has been initiated in Zhu et al. (2022); Liu and Loh (2022); for problems other than density estimation.

## 2.1 Background: Density Estimation and Minimax Rates of Convergence

Without contaminations, the problem of density estimation has been widely studied in non-parametric statistics, see e.g., Stone (1984); Tsybakov (2009). It is well known that some additional smoothness or structural condition must hold in order for density estimation to be possible. If the density is completely unconstrained, then it can wiggle or oscillate so quickly that there is no hope of estimating its value at any given point.

**Hölder smoothness.** One of the most widely studied smoothness conditions is Hölder smoothness. This requires that a function has sufficiently many derivatives, which themselves have to be appropriately smooth. Consider some smoothness parameter  $s > 0$ , and let  $l$  be the largest non-negative integer that is strictly smaller than  $s$ . A one-dimensional density is  $s$ -Hölder smooth if it has an  $l$ th derivative  $f^{(l)}$ , and there is  $L > 0$  (sometimes referred to as the Lipschitz constant) such that

$$|f^{(l)}(x) - f^{(l)}(y)| \leq L|x - y|^{s-l} \text{ for all } x, y \in \mathbb{R}.$$

For instance, if  $s < 1$ , then  $l = 0$  and it is required that

$$|f(x) - f(y)| \leq L|x - y|^s \text{ for all } x, y \in \mathbb{R}.$$

This is interpreted as the ‘‘singularities’’ of  $f$  being behaving at most as  $x \mapsto x^s$ , e.g., for  $s = 1/2$  at most square-root singularities.

In the  $p$ -dimensional case, it is required that all partial derivatives up to order  $l$  exist and satisfy the analogous condition with  $\|x - y\|_2$  instead of  $|x - y|$ .

**Minimax risk.** A fundamental way to evaluate performance in mathematical statistics and statistical decision theory is the minimax risk (Wald, 1949). This is the best performance that any method can achieve in the worst case over a specified problem class. In our setting, it is the best possible squared estimation error

$$\mathbb{E}_{X'_1, \dots, X'_n \sim \tilde{f}}(\hat{f}(x_0; X'_1, \dots, X'_n) - f(x_0))^2$$

that an estimator  $\hat{f}(x_0; X'_1, \dots, X'_n)$  constructed using the contaminated sample  $X'_1, \dots, X'_n$  can achieve, in the worst case over all  $s$ -Hölder smooth densities with a given Lipschitz constant  $L$  and all i.i.d  $W_{q,r}$  perturbations  $\tilde{f}$  of size at most  $\varepsilon$  of the density  $f$ . To be fully precise, the perturbations are also required to be smooth for similar reasons as in the unperturbed case, however, they can have a potentially different Hölder smoothness constant and Lipschitz constant.<sup>4</sup>

**Classical rate of convergence.** One of the cornerstones of nonparametric statistics is that without perturbations, the minimax optimal rate of convergence is (see e.g., Stone, 1984; Tsybakov, 2009, etc)

$$n^{-\frac{2s}{2s+p}}. \quad (1)$$

Moreover, this rate is achieved by kernel density estimation with a suitably smooth and bounded kernel and bandwidth  $h^* \asymp n^{-\frac{1}{2s+p}}$ .

## 2.2 The New Result: Optimal Density Estimation under Wasserstein Contamination

In this context, our new result on optimal density estimation under Wasserstein contaminations shows that the rate of the minimax risk is determined by the maximum of two terms: the standard rate and a term dependent on  $\varepsilon$ .

**Theorem 2.1** (Density estimation under Wasserstein contamination, Chao and Dobriban (2023)). *Let  $p \in \mathbb{N}$ ,  $s > 0$ ,  $L > 0$ ,  $q \in [1, \infty]$ ,  $r \in [1, \infty)$ . Then for all  $n \geq 1$  and  $\varepsilon \in (0, 1]$ , the minimax risk for  $s$ -Hölder smooth density estimation in  $p$  dimensions from  $\varepsilon$ -bounded  $W_{q,r}$  i.i.d. Wasserstein contaminated data has the rate*

$$\max \left\{ n^{-\frac{2s}{2s+p}}, \varepsilon^{\frac{2s}{s+1+p/r}} \right\}.$$

**Optimal estimator.** Moreover, the upper bound is achieved by kernel density estimation with suitably smooth and bounded kernels and bandwidth  $h^* \asymp \max \left\{ n^{-\frac{1}{2s+p}}, \varepsilon^{\frac{1}{s+1+p/r}} \right\}$ . This shows that, compared to the unperturbed case, the minimax risk and the required level of smoothing in the estimator does not change if  $\varepsilon$  is sufficiently small that  $\varepsilon \lesssim n^{-\frac{s+1+p/r}{2s+p}}$ . However, the risk is dominated by the contaminations above that level, and the required level of smoothing is also determined by the magnitude of contaminations.

**Contrast with robustness in the classical Huber  $\varepsilon$ -contamination model.** To put things in context, one key reason why this result is interesting is that it shows that for the Wasserstein-type contaminations, classical methods based on local smoothing can be optimal in a specific problem. This is in stark contrast with the more classically studied Huber  $\varepsilon$  contamination model, where standard smoothing or averaging-based estimators can be very sensitive, and other estimators, such as median-type methods, are required. Thus, complementing also the work of Zhu et al. (2022); Liu and Loh (2022), our results provide a different perspective on robust statistical estimation.

## 3 The Proof: A Human-AI Interaction Workflow

In this section, we explain our workflow for obtaining the proof. The full proof itself can be found in our preprint (Chao and Dobriban, 2023, arxiv v3), and we will not repeat it here. Instead, we will explain the context, our workflow emphasizes the complementary contributions

---

<sup>4</sup>However, it turns out that the rate of convergence as a function of  $n$  and  $\varepsilon$  does not depend on these, so they will not be mentioned further.

of human work and AI, aiming to distill generalizable principles that could be useful more broadly.

### 3.1 Context and Background for the Problem

**Robust statistics.** Our problem can be viewed as a natural question in robust statistics. Contaminations and robust statistics are well-studied, with a large literature focusing on the Huber  $\varepsilon$ -contamination model, see e.g., Huber (1964, 1965, 2004); Hampel et al. (2005), etc. Given a distribution  $P$ , the data are sampled from a distribution  $Q = (1 - \varepsilon)P + \varepsilon P'$ , where  $P'$  is unconstrained in the standard formulation.

**Optimal transport.** Optimal transport has registered great progress over the last twenty years (Villani, 2003, 2008; Ambrosio et al., 2008; Santambrogio, 2015; Figalli and Glaudo, 2021; Peyré and Cuturi, 2019). Similarly, Wasserstein distributionally robust estimation and contamination have become widely studied, see e.g., Esfahani and Kuhn (2015); Lee and Raginsky (2017); Blanchet and Kang (2017); Singh and Póczos (2018); Manole et al. (2022); Weed and Berthet (2019); Blanchet et al. (2019); Chen et al. (2020); Staib (2020); Shafieezadeh Abadeh (2020); Hütter and Rigollet (2021); Duchi et al. (2021). These works typically focus on solving numerical minimax optimization problems algorithmically, where estimators minimize the least favorable risk over a class of contaminations.

**Prior work on Wasserstein contaminations.** In contrast, we aim to investigate robust estimation problems. Prior work by Zhu et al. (2022); Liu and Loh (2022) has initiated this line of inquiry, obtaining a variety of intriguing results, for problems ranging from mean estimation and linear regression to covariance matrix estimation. Their works show that various estimators, such as Wasserstein GANs, can be nearly minimax optimal.

Inspired by these works, starting around February 2021, we started investigating the problem of robustness to Wasserstein distribution shifts, with initial results reported in August 2023 (Chao and Dobriban, 2023, arxiv v1). While some of our results concern mean estimation and linear regression as studied in Zhu et al. (2022); Liu and Loh (2022), here we would like to focus on the problem of nonparametric estimation under Wasserstein distribution shift, which, to the best of our knowledge has not been studied before our work.

### 3.2 Our Initial, Human-only Results

Let  $\mathcal{M}(\varepsilon, s, p, q, r)$  be the minimax risk for  $p$ -dimensional  $s$ -Hölder smooth density estimation under  $\ell_q^r$  Wasserstein perturbations of size  $\varepsilon$ , as defined above. In the initial version of our manuscript (Chao and Dobriban, 2023, arxiv v1), we have considered the special case of  $p = 1$ -dimensional estimation, under  $\ell_2^2$  perturbations, so that  $q = r = 2$ . We have obtained the following result:

$$n^{-2s/(2s+1)} \vee \varepsilon^{4s/(2s+1)} \lesssim \mathcal{M}(\varepsilon, s, 1, 2, 2) \lesssim n^{-2s/(2s+1)} \vee \varepsilon^{2s/(s+2)}.$$

In this result, the term  $n^{-2s/(2s+1)}$  represents the classical minimax optimal rate from (1); and it is quite clear that this represents the correct dependence on the sample size  $n$ . However, the dependence on  $\varepsilon$  is much less clear. Our results above do not match, and the lower bound can be much smaller in  $\varepsilon$  than the upper bound.

There are clear proof strategies for both the upper and the lower bound, which are well-known in the literature.

1. For the **upper bound**, a classical approach (see e.g., Stone, 1984; Tsybakov, 2009, etc) characterizes the bias and the variance of a kernel density estimator.

- For the **lower bound**, there are a variety of classical strategies, such as the Le Cam, Fano, and Assouad bounds, as well as the modulus of continuity (see e.g., Stone, 1984; Tsybakov, 2009; Donoho and Liu, 1991, etc).

In our initial work (Chao and Dobriban, 2023, arxiv v1), we have started by attempting to adapt these strategies. Specifically:

- For the **upper bound**, the same approach applies, but there is also a new contamination bias term. Let  $K : \mathbb{R}^p \rightarrow \mathbb{R}$  be a kernel, and let  $h > 0$  be a bandwidth. Then this contamination bias term equals

$$\frac{1}{h^p} \mathbb{E}_{X'_1 \sim \tilde{f}} \left[ K \left( \frac{X'_1 - x_0}{h} \right) \right] - \frac{1}{h^p} \mathbb{E}_{X_1 \sim f} \left[ K \left( \frac{X_1 - x_0}{h} \right) \right]. \quad (2)$$

In our work, we have used a Lipschitz property for  $K$  and Jensen's inequality to bound this term, obtaining a bias bound of the order of  $\varepsilon/h^2$ .

- In our **lower bound**, we have adapted the modulus of continuity approach, by starting with a normal density  $f = \varphi_\sigma$  with sufficiently small variance  $\sigma^2 > 0$ , and perturbing it with a mean zero function to obtain a contaminated density  $\tilde{f}$ . Then, we upper bounded  $W_{2,2}(f, \tilde{f}) \leq \varepsilon$  (via the Talagrand transportation cost inequality, see Talagrand (1996); Otto and Villani (2000)) and lower bounded  $|f(x_0) - \tilde{f}(x_0)|$ , which—by the modulus of continuity approach—lead to a lower bound on the estimation error.

However, our two bounds did not match. Moreover, despite familiarity with basic techniques and results in optimal transport, we were unable to come up with better results. Further, we were even unsure whether the lower or the upper bounds need improvement, and simulation results were not too helpful, as the problem space is quite large (many possible base densities, perturbations, and estimators). Therefore, we were quite stuck.

### 3.3 Optimal Results via Human-AI Collaboration

For the revision of the paper, starting in October 2025, the present author has decided to try using LLMs for assistance. LLM capabilities had been increasing steadily, and there were already a number of examples of LLMs being helpful with quite advanced mathematical results, as described above. This provided ample motivation.

**Workflow.** The workflow was simple. We copy-pasted the relevant parts of our LaTeX manuscript (problem statement and proof) into<sup>5</sup> GPT-5 via the web interface, and appended a simple prompt such as “I’m working on this theorem. Can you come up with a better upper bound? Please prepare ready to copy-paste fully rigorous LaTeX”. Then, we would send the prompt to GPT-5 Pro, copy-paste the output into a LaTeX editor, and carefully verify the results through our own human expertise. If we identified any mistakes, we simply explained them to the model, and continued the conversation.

The extended test-time compute of GPT-5 Pro was essential, and reasoning took anywhere from a few minutes to 25 minutes per turn. Answers without test-time reasoning were significantly less useful.

**The contributions of the AI.** GPT helped by providing ideas and suggestions on improving both the upper and the lower bounds. In both cases, the argument started with the human-provided reasoning, and improved/sharpened the results. We explain this below.

---

<sup>5</sup>It is possible that other models could work well; we have simply not tried these models in the current work.

## GPT Suggests Improved Bias Bound

User: [Prompt includes LaTeX draft of paper and a query: "Can you double-check this proof from beginning to end, and make sure it is fully correct? Please identify any place where there are incorrect arguments."]

GPT: [...]

The correct formulation is as follows. Define the interpolation

$$T_t(x) = (1-t)x + tT(x), \quad X_t = T_t(X),$$

where  $X \sim P$  and  $T$  is the optimal transport map from  $P$  to  $Q$ . Then

$$\frac{d}{dt} \mathbb{E}[\varphi(X_t)] = \mathbb{E}[\varphi'(X_t) \cdot (T(X) - X)],$$

where the expectation is with respect to the original law of  $X \sim P$ . [...]

**Consequence.** Once the derivative is written correctly, the displayed Cauchy–Schwarz step goes through. The correct bound is

$$|\mathbb{E}_Q[\varphi] - \mathbb{E}_P[\varphi]| \leq \left( \int_0^1 \mathbb{E}_P[\varphi'(X_t)^2] dt \right)^{1/2} W_2(P, Q),$$

which coincides with your statement after noting that  $\mathbb{E}_P[g(X_t)] = \mathbb{E}_{P_t}[g]$  for any test function  $g$ , where  $P_t$  is the law of  $X_t$ .

- For the **upper bound, the key suggestion was a better control of the contamination bias** term from (2). The first observation was to leverage that for our problem, and for<sup>6</sup>  $q \in (1, \infty)$ ,  $r > 1$ , there is always a deterministic optimal transport map  $T$ , as opposed to a generic probabilistic optimal coupling, see e.g., Gangbo and McCann (1996), or Theorem 10.28 or 10.38, and surrounding discussion in Villani (2008). We were aware of this result, but did not consider using it here.

Write  $\varphi(x) = K_h(x_0 - x) = K[(x_0 - x)/h]/h^p$ ,  $\Delta(y) = T(y) - y$ . For  $t \in [0, 1]$ , let  $S_t(y) = (1-t)y + tT(y)$  being the linear interpolation of the identity map with the optimal transport map. **GPT suggested a sharper analysis of the contamination bias**  $(K_h * \tilde{f})(x_0) - (K_h * f)(x_0)$  by writing it leveraging the fundamental theorem of calculus through  $\varphi(T(y)) - \varphi(y) = \int_0^1 \nabla \varphi(S_t(y))^\top \Delta(y) dt$  as

$$\int (\varphi(T(y)) - \varphi(y)) f(y) dy = \int_0^1 \int \nabla \varphi(S_t(y))^\top \Delta(y) f(y) dy dt.$$

**Consequence.** Then, one can use Hölder's inequality to bound the inner product. Let  $\nu_t := (S_t)_\#(fdy)$  be the pushforward measure and let its density be  $\rho_t$ . Let  $r^* > 1$  be the conjugate of  $r$ , defined such that  $1/r^* + 1/r = 1$ , and let  $q^*$  be defined analogously. For each  $t \in (0, 1)$ , we then have

$$\int |\nabla \varphi(S_t(y))^\top \Delta(y)| f(y) dy \leq \left( \int \|\nabla \varphi(z)\|_{q^*}^{r^*} \rho_t(z) dz \right)^{1/r^*} \left( \int \|\Delta(y)\|_q^r f(y) dy \right)^{1/r}.$$

If the density  $\rho_t$  is bounded, then the first term can be bounded by a standard integral in terms of  $\varphi$  against the Lebesgue measure, which depends only on the kernel  $K$  (which

---

<sup>6</sup>The remaining values of  $q$  can be handled by simple limiting arguments, and  $r = 1$  can be handled by Kantorovich duality.

is fixed), and the bandwidth  $h$  (which can be controlled). Moreover, the second term is precisely  $W_{q,r}(\tilde{f}, f) \leq \varepsilon$  and so can be controlled.

This argument leads to the correct upper bound.

**Transcript and extract.** We provide a link to the first transcript where GPT made such a suggestion in a correct form.<sup>7</sup> An extract from our interaction is also provided.<sup>8</sup>

### GPT-5 Suggests Dynamic Benamou-Brenier for Improved Lower Bound (Extract)

User: [Prompt includes LaTeX draft of paper and a query: "Can you go through this proof and see if you can improve either the upper bound and the lower bound until they match?"]

GPT: [...]

The root cause is that Lemma [...] scales like  $W_2^2 \lesssim D^2 \int g_+$ , i.e. it is linear in the bump amplitude  $\tau$  while the correct small-amplitude scaling for  $W_2$  in one dimension is quadratic in  $\tau$ .

[...]

We state and prove the replacement for Lemma [...].

**Lemma 3.1** (Refined local  $W_2$  bound via the dynamic formulation). *Let  $q$  be a probability density on  $\mathbb{R}$  and let  $I \subset \mathbb{R}$  be a finite interval. Let  $g \in L^1(\mathbb{R})$  satisfy  $\text{supp}(g) \subset I$  and  $\int g = 0$ . Write  $G(x) = \int_{-\infty}^x g(t) dt$ . Suppose there exists  $c_I > 0$  such that  $q(x) \geq c_I$  on  $I$  and  $|g|_\infty \leq c_I/2$ . Then  $q \pm g$  are densities and*

$$W_2^2(q + g, q) \leq 2c_I \int_{\mathbb{R}} G(x)^2 dx, \quad W_2^2(q - g, q) \leq 2c_I \int_{\mathbb{R}} G(x)^2 dx.$$

*Proof.* Consider the path of densities  $\rho_t = q + tg$  for  $t \in [0, 1]$ . On  $I$  we have  $\rho_t \geq q - |g|_\infty \geq c_I/2$ , while off  $I$  one has  $\rho_t = q > 0$ . Define  $v_t(x) = -G(x)/\rho_t(x)$ . Then  $\rho_t v_t = -G$  is independent of  $t$ , hence  $\partial_x(\rho_t v_t) = -\partial_x G = -g$  and  $\partial_t \rho_t = g$ ; thus  $\partial_t \rho_t + \partial_x(\rho_t v_t) = 0$ . Applying the Benamou–Brenier dynamic formulation gives

$$W_2^2(q + g, q) \leq \int_0^1 \int_{\mathbb{R}} |v_t(x)|^2 \rho_t(x) dx dt = \int_0^1 \int_{\mathbb{R}} G(x)^2 dx dt = \int_{\mathbb{R}} G(x)^2 dx.$$

Since  $\rho_t(x) \geq c_I/2$  on  $I$  and  $\rho_t = q$  off  $I$ , the bound sharpens to

$$W_2^2(q + g, q) \leq 2c_I \int_{\mathbb{R}} G(x)^2 dx,$$

and the same estimate holds for  $W_2(q - g, q)$ . □

[...]

2. For the lower bound, the key suggestion was to provide an improved bound on the magnitude of the perturbation  $W_{q,r}(f, \tilde{f})$ , via the dynamic Benamou–Brenier formulation of optimal transport (Brenier, 2004).

<sup>7</sup><https://chatgpt.com/share/6922876a-7988-8007-9c62-5f71772af6aa>

<sup>8</sup>Here, GPT actually suggests a fix to a previous, incorrect dynamic formulation. Note that this is for a special case for our problem (one dimension,  $q = r = 2$ ), because it occurred prior to generalizing to a higher dimensional case, etc.

Specifically, GPT made suggestions that we eventually refined into the following lemma:

**Lemma 3.2** (Local  $W_{q,r}$  bound in  $\mathbb{R}^p$  via the dynamic formulation). *Let  $\mathcal{G}$  be a probability density on  $\mathbb{R}^p$ , let  $I = \prod_{j=1}^p [a_j, b_j] \subset \mathbb{R}^p$  be a bounded box, and let  $g \in L^1(\mathbb{R}^p)$  satisfy  $\text{supp}(g) \subset I$  and  $\int_{\mathbb{R}^p} g = 0$ . Assume  $\mathcal{G}(x) \geq c_I > 0$  on  $I$  and  $\|g\|_\infty \leq c_I/2$ . For  $r \in [1, \infty)$ , consider a vector field  $U \in L^r(\mathbb{R}^p; \mathbb{R}^p)$  with  $\text{supp}(U) \subset I$  and  $\nabla \cdot U = g$ . Then*

$$W_{q,r}^r(\mathcal{G} \pm g, \mathcal{G}) \leq \frac{2^{r-1}}{c_I^{r-1}} \int_{\mathbb{R}^p} \|U(x)\|_q^r dx.$$

To prove this statement GPT suggested using the dynamic Benamou-Brenier formulation. This shows that the optimal transport cost is a lower bound on the transportation cost along any path that satisfies a continuity equation.

**Suggested argument.** GPT suggested the following argument to prove that this holds: For  $t \in [0, 1]$  define  $\rho_t := \mathcal{G} + tg$  and  $v_t(x) := -U(x)/\rho_t(x)$ . Since  $\rho_t(x) \geq c_I/2$  on  $I$  and  $U$  vanishes outside  $I$ , this is well defined. We claim that  $(\rho_t, v_t)_{t \in [0,1]}$  satisfies the continuity equation  $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$  with endpoints  $\rho_0 = \mathcal{G}$  and  $\rho_1 = \mathcal{G} + g$ . Indeed  $\partial_t \rho_t = g$ , while  $\nabla \cdot (\rho_t v_t) = \nabla \cdot (-U) = -\nabla \cdot U = -g$ , hence the sum vanishes.

The results of Brenier (2004) then imply the lemma. Using the lemma in our theorem, and verifying the appropriate regularity conditions (also with help from GPT), lead to the desired lower bound.

**Transcript and extract.** We provide a link to the first transcript where GPT made such a suggestion.<sup>9</sup> An extract from our interaction is also provided.<sup>10</sup>

3. Beyond these concrete suggestions, we have also benefited from using GPT in other ways, which are now perhaps relatively well-established, and thus we will not dwell on them too much. For instance, as it is now well-established in the literature (see e.g., Zheng et al., 2023, etc), **we have asked the LLM to be a verifier by checking specific proofs** or writing, and to prepare a list of potential issues. These were a concern for our preprint because we went over several revisions and extensions, which could easily introduce inconsistencies.

This was helpful for catching a number of small notational inconsistencies in our paper (e.g., using both  $p$  and  $d$  for the dimension). The helpfulness stems from being able to human-verify the AI claims (clearly true e.g., for the notation issues) much more quickly than it would take to find them.

**We think that the contributions of the AI were remarkable, and would certainly describe them as “creative” if they had been suggested by a human.** After having worked on this problem for more than two years, and not having solved it, we can certainly appreciate the contribution. The contribution of the AI was crucial, and we estimate that it has saved us up to several months of work. In order to recover the same proofs, we would have essentially needed to read the literature, learn about all the tools, and realize how they can be used in our problem.

**Isn’t this all standard?** It is likely that other researchers, with a deeper knowledge of optimal transport, will view these techniques as “standard”. In a sense, this note reveals the limitations of the author’s knowledge and expertise more than anything. However, we still believe that the contributions of AI are valuable, and that the setting that we described is somewhat typical of individual researchers’ workflow.

---

<sup>9</sup><https://chatgpt.com/share/69227dd1-a00c-8007-8072-26b6e97a0820>

<sup>10</sup>This is also for a special case for our problem (one dimension,  $q = r = 2$ ), because it occurred prior to generalizations. However, this was where the AI gave the initial idea.

1. First of all there are few researchers who have a deep knowledge of each of optimal transport, nonparametric density estimation, and robust statistics. There are even fewer who have showed an active interest in the intersection of the three in terms of published research (we are not aware of any!), and who have a ready availability so that we could have asked for help or guidance. See Figure 2 for a cartoon illustration.
2. Second, in the mathematical sciences, it is quite typical to work in small teams, with up to four-five collaborators. Thus, necessarily, the expertise of most teams is limited.

In contrast, a **key advantage of AI as a collaborator is that it has a vast knowledge, including of the three areas above, and that it is immediately available for consultation.** The LLM has read a vast number of papers and books, and so can make connections between similar problems, results, and techniques, across a wide range. To put it another way, key advantages are its large memory, and its ready anytime availability.

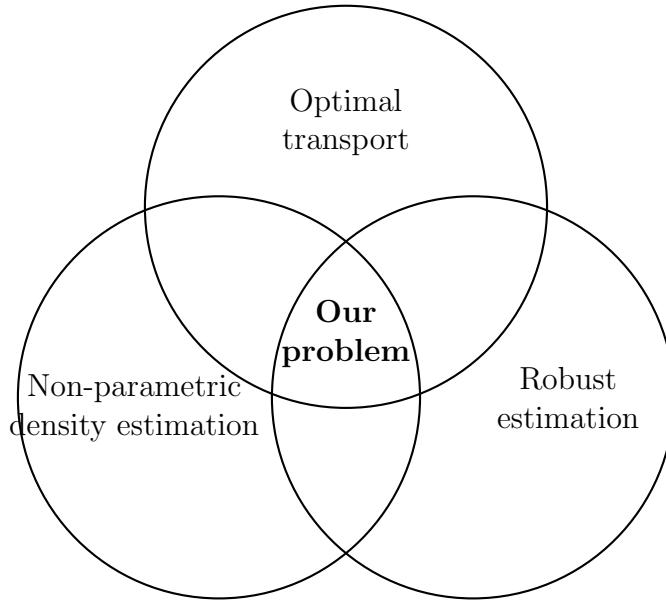


Figure 2: Schematic of the problem setting at the intersection of nonparametric density estimation, robust estimation, and optimal transport.

### 3.4 Limitations of the AI

While the AI clearly made crucial contributions, it was definitely not perfect. It hallucinated references, it glossed over important details, and skipped key steps in the argument. Below, we outline some of these issues, and how we eventually overcame them.

1. For the **upper bound**, there were two major issues.
  - (a) A crucial component is **the existence of the optimal transport map  $T$** . Initially, the AI suggested that this follows from certain results in the book by Santambrogio (2015). After carefully reading those results, we realized that they were only for the Euclidean case where  $q = 2$ . After re-prompting the AI pointing this out, it suggested Chapter 10 of the book by Villani (2008), which however, was quite hard to parse, due to tackling a far more general setting than we needed. Eventually, after additional AI-assisted search, we tracked down the reference by Gangbo and McCann (1996), a standard in the field. This shows that, **even if a**

**general suggestion is in the right direction, many details can be missing** and may need to be filled in by additional work. Ultimately, our human expertise was necessary to verify the correctness.

- (b) Second, another crucial component of the argument is **the boundedness of the density  $\rho_t$** . For this, the AI suggested another result from Santambrogio (2015), but that again only covered the Euclidean case.

Tracking down a correct reference proved tricky. Similar results are well-known in the literature, but most of them for slightly different assumptions; for instance Proposition 7.29 of Santambrogio (2015) is for Euclidean (as opposed to  $\ell_q$ ) norms, while the results of Ohta (2009) are stated for squared (as opposed to power- $r$ ) distances.

**The eventual solution.** After unsuccessfully querying the AI multiple times to generalize these results, and to find the required results in the literature, we were eventually successful by providing a sketch of the proof that the AI completed. Specifically, we realized that, as in the Riemannian Borell-Brascamp-Lieb inequality (Cordero-Erausquin et al., 2001) and its adaptation to Finsler geometry (Ohta, 2009), the determinant of the Jacobian of the transport map can be bounded along the interpolation path. In fact, given the properties of the optimal transport map (Gangbo and McCann, 1996), the proof was convenient.

The proof of this part was highly iterative and interactive, requiring many steps of careful reading of the literature, prompting the AI, and verifying its solutions. There were many mis-steps, but they were all eventually resolved. This was the toughest part of the argument, where we thought that we hit a wall several times.

2. For the **lower bound**, the AI initially suggested to use the dynamic Benamou-Brenier formulation from the book by Santambrogio (2015). However, as we have already mentioned, that only covers the Euclidean case. After the AI has not been able to find a correct reference even after multiple attempts, we were eventually led to check out the original paper by Brenier (2004). Fortunately, their results were enough to cover our case.
3. There were also a number of claims that we tried to prove that the AI struggled with. A priori, it was hard to tell whether or not it could work. **Regardless of whether or not the AI found a correct proof of a claim, it would almost always claim that it had succeeded.** This compelled us to verify its proof, which required time and effort. Mistakes were often subtle, requiring careful attentive reading to find.

**When we found and reported a mistake, the AI would typically generate a longer and more complicated purported fix.** The cycle could repeat many times over. Understandably, the AI is trained never to give up; which is often desirable, but sometimes giving up is better than running in circles forever. Our overall work required hundreds of queries. At some point, we had to use our best judgment to stop, give up, and exit.

Overall, we hope that the above discussion provides a balanced overview of the capabilities and limitations of using current publicly available AI systems for math, and what one may expect when using them.

## 4 Commentary and Discussion

We feel that several important comments must be made or repeated, and discussion points must be raised, in order to have a fully balanced report:

- 1. Human expertise was indispensable, and the role of education was and remains crucial.** For some background, over the last 15 years, the present author has acquired a good amount of experience with research in the mathematical sciences (reading and writing proofs, literature search), writing papers, and a good working knowledge of basic optimal transport, at the level of Figalli and Glaudo (2021). These skills took many years to develop, starting from the school level, through doctoral training, and continuing at the level of a faculty member.

Without such skills, we feel that it would be effectively impossible to properly use current AI tools for advanced mathematical research. Therefore, **current AI tools do not diminish in the slightest the importance of education.** Students should still be educated with the same basic mathematical skills. Even more, students with more advanced mathematical skills are likely to be able to leverage AI to much greater impact; hence the role of education becomes even more important.

Instead, the implication on education is that, possibly, if one masters the basics of certain advanced fields or topics (e.g., the basics of existence theorems for optimal transport maps), one may be able to use AI as an aid to work at a more sophisticated level, tackling harder problems faster, than otherwise possible. Moreover, it is also possible that some level of training in the use of AI tools could help.

- 2. The feeling of AI-empowerment.** Mathematical research requires coming up with creative ideas, but can also involve significant amounts of intellectual effort on relatively routine but lengthy calculations (think of integrals requiring pages of computations); which are few mathematicians' favorite work. Moreover, these calculations are often highly customized, and so the only help/automation that one could hope for is to ask a good collaborator or student to work on them.

Since AI can excel at such calculations, it can greatly reduce the amount of required exhausting work. Of course, AI outputs need to be checked, but this is often less challenging than performing the calculations. This can be highly empowering. By saving energy from routine calculations, mathematicians can focus more on creative ideas, leading to an increase in both productivity and well-being.

To summarize, in this note we have documented how an instance of human-AI collaboration led to the solution of a research problem in mathematical statistics. While current AI systems were not perfect, they already provided valuable suggestions and greatly sped up the research process.

## References

- L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- J. Blanchet and Y. Kang. Distributionally robust groupwise regularization estimator. In *Asian Conference on Machine Learning*. PMLR, 2017.
- J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.
- Y. Brenier. Extended monge-kantorovich theory. In *Optimal Transportation and Applications: Lectures given at the CIME Summer School, held in Martina Franca, Italy, September 2-8, 2001*, pages 91–121. Springer, 2004.
- S. Bubeck, C. Coester, R. Eldan, T. Gowers, Y. T. Lee, A. Lupsasca, M. Sawhney, R. Scherrer, M. Sellke, B. K. Spears, D. Unutmaz, K. Weil, S. Yin, and N. Zhivotovskiy. Early science acceleration experiments with gpt-5, 2025. URL <https://arxiv.org/abs/2511.16072>.

- P. Chao and E. Dobriban. Statistical estimation under distribution shift: Wasserstein perturbations and minimax theory. *arXiv preprint arXiv:2308.01853*, 2023.
- R. Chen, I. C. Paschalidis, et al. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.
- D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger. A riemannian interpolation inequality à la borell, brascamp and lieb. *Inventiones mathematicae*, 146(2):219–257, 2001.
- D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, pages 668–701, 1991.
- J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.
- P. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171, 05 2015.
- M. Feldman and A. Karbasi. G\”odel test: Can large language models solve easy conjectures? *arXiv preprint arXiv:2509.18383*, 2025.
- A. Figalli and F. Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. EMS, 2021.
- W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996. doi: 10.1007/BF02392620.
- F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics*. Wiley, 2005.
- P. J. Huber. *Robust statistics*. John Wiley & Sons, 2004.
- P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.
- P. J. Huber. A Robust Version of the Probability Ratio Test. *The Annals of Mathematical Statistics*, 36(6):1753 – 1758, 1965.
- J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.
- U. Jang and E. K. Ryu. Point convergence of nesterov’s accelerated gradient method: An ai-assisted proof. *arXiv preprint arXiv:2510.23513*, 2025.
- M. I. Jordan. A message from the president: What are the open problems in bayesian statistics? *ISBA Bulletin*, 18(1):1–4, 2011.
- J. Lee and M. Raginsky. Minimax statistical learning with wasserstein distances. In *Neural Information Processing Systems*, 2017.
- Z. Liu and P.-L. Loh. Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference: A Journal of the IMA*, 12(1):312–362, 2022.
- T. Manole, S. Balakrishnan, and L. Wasserman. Minimax confidence intervals for the Sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252 – 2345, 2022.
- S.-i. Ohta. Finsler interpolation inequalities. *Calculus of Variations and Partial Differential Equations*, 36(2):211–249, 2009.
- F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.
- G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- A. Salim. Accelerating mathematical research with language models: A case study of an interaction with gpt-5-pro on a convex analysis problem. *arXiv preprint arXiv:2510.26647*, 2025.
- F. Santambrogio. Optimal transport for applied mathematicians. *Birkhäuser, NY*, 55(58-63): 94, 2015.

- S. Shafieezadeh Abadeh. Wasserstein distributionally robust learning. Technical report, EPFL, 2020.
- S. Singh and B. Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.
- M. J. Staib. *Learning and optimization in the face of data perturbations*. PhD thesis, Massachusetts Institute of Technology, 2020.
- C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pages 1285–1297, 1984.
- M. Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.
- C. Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003.
- C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
- A. Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205, 1949.
- L. Wasserman. The biggest unsolved problem. Blog post on \*Normal Deviate\*, June 2012. URL <https://normaldeviate.wordpress.com/2012/06/21/90/>. Posted June 21, 2012.
- J. Weed and Q. Berthet. Estimation of smooth densities in wasserstein distance. In *Conference on Learning Theory*, pages 3118–3119. PMLR, 2019.
- L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- B. Zhu, J. Jiao, and J. Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256 – 2283, 2022.