Marginal-Fair Aggregation for Distributed Mixture of Experts via Optimal Transport

Anonymous ACL submission

Abstract

Distributed Mixture-of-Experts (MoE) models are increasingly trained on decentralized, heterogeneous data, requiring server-side procedures that can synthesize a single global MoE from many locally trained ones. Recent optimal transport (OT) aggregators reduce the $M \times K$ collection of local experts to a K-component global model by minimizing an average transportation divergence to local MoEs. While consistent, this average objective is fundamentally biased: clients with more data, sharper expert distributions, or richer routing patterns disproportionately determine the global geometry, leaving low-resource or atypical clients systematically misrepresented. We propose Marginal-Fair OT-MoE (MFOT-MoE), an aggregation framework that replaces the weighted-sum OT objective with a minimax marginal-fair objective: the global MoE minimizes the worst-case transportation divergence to any client's expert family, equidistancing the aggregator from all local marginals. We connect this objective to the recently proposed Marginal-Fair Sliced Wasserstein Barycenter (MFSWB) and adapt its three surrogate variants (s-, us-, es-MFSWB) to the conditional MoE setting. The resulting algorithm modifies only the server-side reduction step of standard OT-MoE: an additional softmax over per-client transport costs reweights local experts during the M-step, with no client-side or communication change. We prove well-posedness, approximate monotonic descent of an adapted majorization-minimization solver (with explicit $O(\varepsilon/\tau)$ residual from entropic OT), and a finite-sample fairness generalization bound; we further show that standard OT-MoE arises as the high-temperature limit. Experiments on synthetic Gaussian regression and federated tabular benchmarks with a fairness stress test under quantity skew show that MFOT-MoE matches OT-MoE on utility metrics while reducing the F-metric by up to 85% and worst-client MSE by up to 49%.

1 Introduction

Mixture-of-Experts (MoE) architectures are now central to scalable deep learning, enabling conditional computation through sparse routing over specialized subnetworks [Shazeer et al., 2017]. As model size and data privacy constraints grow, training MoE models on decentralized data has become a practical necessity in domains ranging from multi-institutional medical imaging to large language model fine-tuning over edge devices [McMahan et al., 2017]. In this distributed setting, each of M clients trains a local MoE with K experts on its own data, and a central server is tasked with synthesizing a single global K-expert MoE that summarizes the federation without further data exchange.

Aggregating locally trained MoEs is fundamentally different from aggregating monolithic networks. Naive parameter averaging breaks the gating–expert coupling that defines the mixture, and concatenating all local experts produces an unwieldy $M \times K$-component model that violates the original architectural budget. To resolve this structural mismatch, recent work casts MoE aggregation as a measure-theoretic *reduction* problem solved with optimal transport: the global K-expert MoE is the minimizer of an expected transportation divergence to the flattened collection of $M \times K$ local experts, and a majorization-minimization (MM) solver alternates between transport plans and expert updates [Chamroukhi and Pham, 2023]. This formulation preserves MoE structure, admits a closed-form M-step for Gaussian experts, and has been extended to Bayesian personalization [Wei et al., 2025] and adapter fusion via OT permutation matching.

The structural fairness problem. A subtle but important issue persists in all of these methods. By minimizing a *weighted sum* of OT distances with weights $\lambda_m \propto N_m/N$, the standard reduc-

tion is dominated by clients with larger or richer data. Crucially, this is not merely a tuning choice: even under uniform weights $\lambda_m = 1/M$, clients with higher-entropy gating or larger expert variance have steeper gradients of the OT cost at the optimum, pulling the global barycenter disproportionately toward themselves. The resulting global model performs well on those clients but *systematically degrades on under-represented ones—not because the optimizer fails, but because the objective minimizes expected transport cost rather than worst-case cost.* This is a structural property of the aggregation *objective*, not an artifact of optimization.

Existing fair federated learning methods such as q-FFL [Li et al., 2020] or Agnostic FL [Mohri et al., 2019] reweight clients by scalar loss, which can correct the optimization landscape but leaves the *geometric imbalance in expert distribution space* intact. Even if q-FFL equalizes task losses, the global MoE's expert geometry—which experts activate for which inputs, and what conditional distributions they cover—is still determined by the transport structure of the reduction objective. Routing-level methods such as FairSMoE [Park et al., 2025] address attribute fairness within a single MoE, not aggregation across MoEs.

For instance, consider $M = 3$ hospital sites where site 1 has $10\times$ the data of sites 2–3 with rare subtypes: under OT-MoE the global expert geometry is pulled toward site 1's structure, and gating under-activates the experts needed for sites 2–3. Standard DRO on client losses does not correct this—the loss gap may close while the geometric mismatch persists. What is needed is an aggregation *objective* that is OT-geometric, MoE-structural, and provably equidistant from every client's expert family. The recently introduced Marginal-Fair Sliced Wasserstein Barycenter (MFSWB) [Nguyen et al., 2025] offers exactly such a primitive in the unconditional measure setting: it replaces the sum of OT distances with a maximum, yielding a barycenter that is geometrically equidistant from all input marginals. Translating MFSWB into the conditional, MoE-structured federated setting is non-trivial because the marginals are not points in $\mathcal{Y}$ but *covariate-conditional mixtures* on $\mathcal{P}(\mathcal{Y})$.

Our contribution. We propose **Marginal-Fair OT-MoE** (MFOT-MoE), the first aggregation framework that combines the MoE-structural reduction of OT-MoE with the geometric fairness of MFSWB. The global MoE is defined as

$$\bar{f}^{R,\mathrm{MF}} = \arg\min_{g \in \mathcal{M}_K} \mathbb{E}_x \left[\max_{m \in [M]} \mathcal{T}_c\Big(\widehat{f}_m, g; x\Big) \right], \tag{1}$$

where $\widehat{f}_m$ is the local MoE conditional density at client m and $\mathcal{T}_c$ is the conditional transportation divergence used in OT-MoE. The minimax structure prevents the aggregator from improving its objective by neglecting any single client. To make the optimization tractable, we adapt three MFSWB surrogates – s-MFSWB, us-MFSWB, and es-MFSWB – to the conditional setting and integrate them into the OT-MoE MM solver via a softmax reweighting of clients. The change is minimal: the server replaces uniform-or-proportional client weighting with $\omega_m \propto \exp(\mathcal{T}_c(\widehat{f}_m, g)/\tau)$ during the M-step. As $\tau \to 0$, MFOT-MoE recovers the hard minimax (1); as $\tau \to \infty$, it recovers standard OT-MoE. In summary:

- We formulate marginal-fair OT aggregation for distributed MoE as a conditional minimax barycenter problem on $\mathcal{P}(\mathcal{Y})$, addressing client-level dominance that is intrinsic to weighted-sum OT aggregation.
- We derive an adapted majorization-minimization algorithm that integrates the three MFSWB surrogate variants into the OT-MoE server step with no additional communication and only an $O(M)$ per-round overhead.
- We establish well-posedness and worst-client optimality of the population objective; approximate monotonic descent of the MM solver under entropic regularization; and a finite-sample fairness generalization bound with explicit τ-dependence. We also prove that uniform-weight OT-MoE arises as the high-temperature limit.
- We provide an evaluation protocol with quantity-skew stress tests, four heterogeneity regimes, surrogate ablations, wall-clock overhead measurements, and sensitivity analyses for ε and S. MFOT-MoE reduces F-metric by 85–90% over OT-MoE while matching utility, with $< 3\%$ per-round overhead for s-MFOT.

2 Related Work

OT-based MoE aggregation. Chamroukhi and Pham [2023] introduced OT-MoE, the first principled framework for reducing an $M \times K$ collection of locally trained experts to a K-component

global MoE via minimization of an expected conditional transportation divergence. The method is consistent and admits a closed-form MM solver for Gaussian experts. Subsequent work has applied Wasserstein barycenters to other federated objects: FedWBA aggregates Bayesian posteriors via particle-based barycenters [Wei et al., 2025]; FedAlign-MoE aligns expert routing via cosine-similarity matching [Fang et al., 2026]; and Singh and Jaggi [2020] fuse independently trained networks via OT-based neuron alignment. None of these methods incorporates a marginal-fairness criterion: they all minimize a weighted sum of transport costs.

Fairness and DRO in federated learning. Client-level fair FL aims for uniform performance across participants via reweighting by inverse-performance (q-FFL, Li et al., 2020), distributionally robust optimization (Agnostic FL, Mohri et al., 2019; FedDRO, Anonymous, 2024), or personalization with proximal regularization (Ditto, Li et al., 2021). DRO-based methods directly address worst-client performance by formulating a min-max problem over client loss distributions, making them natural comparators to our minimax framing. The key distinction is that they operate on scalar client losses in parameter space, while MFOT-MoE minimizes a worst-case *transport divergence* in distribution space—correcting a geometric imbalance that scalar reweighting cannot address.

Fairness in MoE. FairSMoE [Park et al., 2025] mitigates demographic disparities by modifying the routing of a single sparse MoE; Lee and Choi [2026] diagnose routing sensitivity to spurious attributes. These works address fairness *within* a single MoE model and require demographic labels, whereas MFOT-MoE addresses fairness *across* clients during aggregation and requires no auxiliary attribute information.

Marginal-fair Wasserstein barycenters. Nguyen et al. [2025] introduce MFSWB and three computationally tractable surrogates – s-, us-, and es-MFSWB – that minimize the maximum sliced-Wasserstein distance to a set of input marginals, thereby producing geometrically equidistant barycenters. Our work transports this idea from unconditional measures on $\mathbb{R}^d$ to covariate-conditional mixtures on $\mathcal{P}(\mathcal{Y})$ and integrates it into an MM-based MoE aggregation pipeline. We also build on classical foundations of Wasserstein barycenters [Agueh and Carlier, 2011] and the sliced family of distances [Bonneel et al., 2015, Kolouri et al., 2019].

Federated MoE under heterogeneity. Recent works study heterogeneous architectures and routing alignment in FL-MoE: HMoE [Zhao et al., 2024] considers heterogeneous expert sizes, Elastic MoE [Čizek et al., 2026] relaxes capacity constraints across clients, and FedAlign-MoE [Fang et al., 2026] proposes cosine-similarity-based expert matching. These methods focus on architectural or routing flexibility, not on the geometric fairness of the aggregation objective.

3 Background and Problem Setup

3.1 Optimal transport and (sliced) Wasserstein barycenters

For probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ with finite p-th moments and a ground cost c, the p-Wasserstein distance is $\mathrm{W}_p(\mu,\nu) = (\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)^p d\pi(x,y))^{1/p}$, where $\Pi(\mu,\nu)$ is the set of couplings [Peyré and Cuturi, 2019]. The *sliced* Wasserstein distance averages 1D projected W_p distances over the unit sphere $\mathbb{S}^{d-1}$ [Bonneel et al., 2015, Kolouri et al., 2019]: $\mathrm{SW}_p^p(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(\theta_\sharp\mu, \theta_\sharp\nu) d\theta$, where $\theta_\sharp$ is the pushforward by $x \mapsto \theta^\top x$.

Given marginals $\{\mu_i\}_{i=1}^N$ and weights $\lambda_i > 0$, the classical Wasserstein barycenter [Agueh and Carlier, 2011] is $\nu^\star = \arg\min_\nu \sum_{i=1}^N \lambda_i \mathrm{W}_p^p(\mu_i,\nu)$. The *marginal-fair sliced Wasserstein barycenter* (MFSWB, Nguyen et al., 2025) replaces the weighted sum with a maximum:

$$\nu^\star_{\mathrm{MF}} = \arg\min_\nu \max_{i\in[N]} \mathrm{SW}_p^p(\mu_i,\nu). \tag{2}$$

Since (2) is non-smooth, three surrogates are introduced: **s-MFSWB** replaces SW_p^p inside the max with a Monte-Carlo estimator and uses a hard maximum; **us-MFSWB** swaps expectation and max, $\min_\nu \mathbb{E}_\theta[\max_i \mathrm{W}_p^p(\theta_\sharp\mu_i, \theta_\sharp\nu)]$, which admits unbiased gradients; **es-MFSWB** draws θ from an energy-based distribution proportional to $\exp(\max_i \mathrm{W}_p^p)$ to focus on directions exposing fairness violations. The chain s $\leq$ us $\leq$ es-MFSWB holds, and us-MFSWB is a lower bound on the multi-marginal Wasserstein distance with maximal ground cost [Nguyen et al., 2025].

3.2 Mixture-of-Experts and OT-MoE aggregation

A conditional MoE on $\mathcal{X} \to \mathcal{P}(\mathcal{Y})$ with K experts is

$$f(y \mid x) = \sum_{k=1}^{K} \pi_k(x; \boldsymbol{\alpha}) \varphi_k(y \mid x; \boldsymbol{\beta}_k), \quad (3)$$

where $\boldsymbol{\pi}(x; \boldsymbol{\alpha}) \in \Delta_{K-1}$ is a softmax gating network and $\varphi_k(\cdot \mid x; \boldsymbol{\beta}_k)$ is the k-th expert distribution [Shazeer et al., 2017]. We write $\mathcal{M}_K$ for the space of K-expert MoE conditional densities.

Consider M clients, each with private data $\mathcal{D}_m$. Client m trains a local MoE $\widehat{f}_m \in \mathcal{M}_K$ with parameters $(\widehat{\boldsymbol{\alpha}}^{(m)}, \{\widehat{\boldsymbol{\beta}}_k^{(m)}\}_{k=1}^K)$ and uploads them to the server. The naive aggregation $\bar{f}^W = \sum_m \lambda_m \widehat{f}_m$ produces an MK-component mixture, breaking the K-expert budget. Chamroukhi and Pham [2023] cast aggregation as the OT *reduction* problem

$$\bar{f}^R = \arg\min_{g \in \mathcal{M}_K} \mathcal{R}_c(\bar{f}^W, g),$$

$$\mathcal{R}_c(\bar{f}^W, g) = \mathbb{E}_{x \sim \mathcal{D}_X}\left[\min_{P \in \mathcal{P}_x(\widehat{\boldsymbol{\pi}}, \boldsymbol{\pi}^g)} \sum_{\ell, j} P_{\ell j} c(\widehat{\varphi}_\ell, \varphi_j^g)\right], \quad (4)$$

where ℓ flattens client-expert pairs (m, k) over $[MK]$, $j \in [K]$, $\widehat{\varphi}_\ell = \widehat{\varphi}_k^{(m)}$, $\boldsymbol{\pi}^g$ is the gating of g, and the row marginals of the transport plan P equal $\widehat{\boldsymbol{\pi}}_\ell = \lambda_m \widehat{\pi}_k^{(m)}(x)$. The cost $c(\cdot, \cdot)$ is a divergence between expert conditionals (e.g., KL for Gaussians). Chamroukhi and Pham [2023] solve (4) via an MM algorithm with a closed-form M-step for Gaussian experts, and prove consistency: $\bar{f}^R \xrightarrow{p} f^\star$ as $N \to \infty$ when each $\widehat{f}_m$ is consistent.

3.3 Why average OT aggregation is unfair

The reduction objective (4) is a *weighted sum* of per-client transportation costs:

$$\mathcal{R}_c(\bar{f}^W, g) = \sum_{m=1}^{M} \lambda_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)], \quad (5)$$

where $\mathcal{T}_c(\widehat{f}_m, g; x)$ denotes the conditional OT cost between $\widehat{f}_m(\cdot \mid x)$ and $g(\cdot \mid x)$ (Lemma A.1, Appendix A). When $\lambda_m \propto N_m$, dominant clients drive the optimum; when $\lambda_m = 1/M$, "richness" (e.g., higher-entropy gating, larger expert variance) still dictates which $\widehat{f}_m$ contributes most to the gradient. Neither choice prevents systematic neglect of distinct, low-resource client distributions. This is a property of the *objective*, not the optimizer.

4 Marginal-Fair OT Aggregation for Distributed MoE

4.1 Minimax marginal-fair objective

We propose the **Marginal-Fair OT-MoE (MFOT-MoE)** reduction:

Definition 4.1 (Marginal-Fair OT-MoE)**.** *Given local MoEs $\{\widehat{f}_m\}_{m=1}^M \subset \mathcal{M}_K$, the marginal-fair aggregated MoE is*

$$\bar{f}^{R,\mathrm{MF}} \in \arg\min_{g \in \mathcal{M}_K} \mathbb{E}_{x \sim \mathcal{D}_X}\left[\max_{m \in [M]} \mathcal{T}_c\Big(\widehat{f}_m, g; x\Big)\right]. \quad (6)$$

At any optimum, no client can be made worse off by the aggregator without making some other client equally bad: the aggregator is *equidistant* from all active client marginals (formalized in Theorem 5.2). This is the conditional, MoE-structured analogue of the s-MFSWB objective (2) with marginals replaced by full client-level conditional MoE densities.

4.2 Three surrogates for the conditional setting

The hard maximum in (6) is non-smooth, making direct gradient-based optimization unstable. We adapt the three MFSWB surrogates of Nguyen et al. [2025] to our conditional setting.

s-MFOT (smoothed max). Replace the max with a log-sum-exp:

$$\mathcal{L}_{\mathrm{s}}^{\tau}(g) = \mathbb{E}_x\left[\tau \log\left(\frac{1}{M}\sum_{m=1}^{M} e^{\frac{1}{\tau}\mathcal{T}_c(\widehat{f}_m, g; x)}\right)\right]. \quad (7)$$

As $\tau \to 0^+$, $\mathcal{L}_{\mathrm{s}}^{\tau} \to$ (6); as $\tau \to \infty$, $\mathcal{L}_{\mathrm{s}}^{\tau} \to \frac{1}{M}\sum_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)] + O(\tau^{-1})$, i.e., uniform-weight OT-MoE.

us-MFOT (unbiased slicing). Use sliced cost components c_θ obtained from 1D projections $\theta \in \mathbb{S}^{d_\mathcal{Y}-1}$, define $\mathcal{T}_{c_\theta}(\widehat{f}_m, g; x)$ as the conditional sliced transport cost, and form

$$\mathcal{L}_{\mathrm{us}}(g) = \mathbb{E}_{x,\theta}\left[\max_{m \in [M]} \mathcal{T}_{c_\theta}(\widehat{f}_m, g; x)\right], \quad (8)$$

with $\theta \sim \mathrm{Unif}(\mathbb{S}^{d_\mathcal{Y}-1})$. Swapping expectation and max yields an unbiased gradient under standard regularity (Theorem 5.3).

es-MFOT (energy slicing). Sample slicing directions from an energy distribution that focuses on fairness-violating projections, $q(\theta) \propto \exp(\max_m \mathcal{T}_{c_\theta}(\widehat{f}_m, g; x))$, sampled via Langevin or self-normalized importance sampling, and define $\mathcal{L}_{\text{es}}$ as the expectation under q. By the importance-weighted argument of Nguyen et al. [2025], $\mathcal{L}_{\text{s}} \leq \mathcal{L}_{\text{us}} \leq \mathcal{L}_{\text{es}}$, with es-MFOT providing the tightest upper bound on the hard minimax.

Remark 4.2 (Role of slicing in the MM algorithm)**.** *For us-/es-MFOT, slicing applies exclusively to the per-client cost $T_m^{(t)}$ in the fairness reweighting (Step 5): the E-step Sinkhorn plans (Step 3) and expert M-step (Step 6) use the original full conditional cost c unchanged. us-MFOT replaces $T_m^{(t)} = \mathbb{E}_x[\mathcal{T}_c]$ with $T_m^{(t)} = \mathbb{E}_{x,\theta}[\mathcal{T}_{c_\theta}]$, averaged over S directions $\theta \sim \text{Unif}(\mathbb{S}^{d_\mathcal{Y}-1})$; es-MFOT uses the energy measure $q(\theta) \propto \exp(\max_m \mathcal{T}_{c_\theta})$ via self-normalized importance weighting. Slicing reduces fairness-weight computation from $O(MK^2/\varepsilon)$ to $O(MSK)$ per round.*

4.3 MM algorithm with marginal-fair reweighting

We adapt the alternating MM framework of Chamroukhi and Pham [2023] by replacing its weighted-sum M-step with a marginal-fair reweighted M-step. At iteration t with current global MoE $g^{(t)}$, we compute per-client transport costs $T_m^{(t)} = \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g^{(t)}; x)]$ and *adaptive client weights*

$$\omega_m^{(t)} = \frac{\exp(T_m^{(t)}/\tau)}{\sum_{m'=1}^{M} \exp(T_{m'}^{(t)}/\tau)}, \tag{9}$$

which are exactly the gradient weights of $\mathcal{L}_{\text{s}}^{\tau}$ with respect to its constituent client costs. We then solve the OT-MoE M-step with $\lambda_m \leftarrow \omega_m^{(t)}$. The full procedure is given in Algorithm 1.

Closed-form M-step for Gaussian experts. When experts are Gaussian linear regressors, $\widehat{\varphi}_k^{(m)}(y \mid x) = \mathcal{N}(y; x^\top \widehat{\boldsymbol{\beta}}_k^{(m)}, \widehat{\sigma}_k^{(m)2})$, with c taken to be squared Wasserstein-2 between Gaussians, the M-step reduces to a weighted least-squares solve with weights $\omega_m^{(t)} P_{kj}^{(m,x)}$. The closed form parallels Proposition 4 of Chamroukhi and Pham [2023], modified by the multiplicative factor $\omega_m^{(t)}$; details in Appendix B. The gate update in Step 7 is deliberately unweighted (uniform client average of transport-plan marginals): injecting $\omega_m^{(t)}$ there would bias routing toward the worst-performing client each round and amplify client-specific routing artefacts. Fairness reweighting thus exclusively governs expert-parameter updates (Step 6).

Algorithm 1 Marginal-Fair OT-MoE Aggregation (MFOT-MM)

Require: Local MoEs $\{\widehat{f}_m\}_{m=1}^M \subset \mathcal{M}_K$; supporting sample $\mathcal{D}_s$; temperature schedule $\{\tau_t\}$; iterations T; surrogate variant $\text{v} \in \{\text{s}, \text{us}, \text{es}\}$; Sinkhorn ε.

1: **Initialize:** $g^{(0)} \in \mathcal{M}_K$ with parameters $(\boldsymbol{\alpha}^{(0)}, \{\boldsymbol{\beta}_j^{(0)}\})$.

2: **for** $t = 0, 1, \ldots, T-1$ **do**

3: **E-step.** For each $x \in \mathcal{D}_s$, m, build $C_{kj}^{(m,x)}$ and solve $P^{(m,x)} = \arg\min_{P \in \Pi} \langle P, C^{(m,x)} \rangle + \varepsilon \mathsf{H}(P)$ via Sinkhorn.

4: **Per-client cost.** $T_m^{(t)} \leftarrow \frac{1}{|\mathcal{D}_s|} \sum_x \langle P^{(m,x)}, C^{(m,x)} \rangle$.

5: **Reweighting.** $\omega_m^{(t)} \leftarrow \text{softmax}(T^{(t)}/\tau_t)_m$ (variant v).

6: **M-step.** For $j \in [K]$: $\boldsymbol{\beta}_j^{(t+1)} = \arg\min_{\boldsymbol{\beta}} \sum_{x,m,k} \omega_m^{(t)} P_{kj}^{(m,x)} c(\widehat{\varphi}_k^{(m)}, \varphi_{\boldsymbol{\beta}})$.

7: **Gate.** Fit $\boldsymbol{\pi}^{(t+1)}(x)$ to $\frac{1}{M} \sum_{m,k} P_{kj}^{(m,x)}$.

8: **Anneal.** $\tau_{t+1} \leftarrow \max(\tau_{\min}, \gamma \tau_t)$.

9: **end for**

10: **Return** $g^{(T)} = (\boldsymbol{\alpha}^{(T)}, \{\boldsymbol{\beta}_j^{(T)}\})$.

Remark 4.3 (Entropic regularization bias)**.** *Algorithm 1 uses Sinkhorn with regularization $\varepsilon > 0$. Since the fairness reweighting $\omega_m^{(t)}$ depends only on* relative *ordering of T_m values, and the entropic bias is shared across all clients and approximately constant in m, the bias partially cancels in $\omega_m^{(t)} \propto \exp((T_m^{(t)} - \bar{T}^{(t)})/\tau)$. The F-metric (which measures pairwise cost differences) is consequently more robust to ε than the absolute utility metric. We use $\varepsilon = 0.1$ throughout.*

Complexity and communication. The only addition over OT-MoE is the per-client softmax reweighting, which costs $O(M)$ per round once $T_m^{(t)}$ are computed. The Sinkhorn step is $O(MK^2/\varepsilon)$. Communication is unchanged: clients still upload only their local MoE parameters once. MFOT-MoE thus retains the single-round federated frugality of OT-MoE [Chamroukhi and Pham, 2023].

5 Theoretical Analysis

We provide three guarantees: well-posedness and worst-client optimality of the population objective; *approximate* monotonic descent of MFOT-MM under entropic regularization, with an explicit $O(\varepsilon/\tau)$

residual; and a finite-sample fairness generalization bound with explicit τ-dependence. We also show that uniform-weight OT-MoE arises as the high-temperature limit. Full proofs appear in Appendix A.

Assumption 5.1 (Regularity)**.** *(i) $\mathcal{Y} \subset \mathbb{R}^{d_\mathcal{Y}}$ is compact; $\mathcal{X}$ is a Polish space. (ii) The cost $c : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$ is continuous and bounded on the relevant domain. (iii) Each $\widehat{\varphi}_k^{(m)}(\cdot \mid x)$ has bounded second moments uniformly in x. (iv) The class $\mathcal{M}_K$ is parameterized by $\Theta \subset \mathbb{R}^p$ with Θ compact and the parameter map $\boldsymbol{\theta} \mapsto g_{\boldsymbol{\theta}}$ continuous in the weak topology. (v) The map $g \mapsto \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)]$ is continuously Fréchet differentiable on the interior of $\mathcal{M}_K$ for each m.*

For Gaussian linear experts with W_2 cost, (i)–(iv) hold under bounded regression coefficients and bounded noise variance, and (v) holds since $\mathbb{E}_x[\mathcal{T}_c]$ is a weighted quadratic form in $\boldsymbol{\theta}$.

Theorem 5.2 (Well-posedness and worst-client optimality)**.** *Under Assumption 5.1, the minimum in (6) is attained on $\mathcal{M}_K$. Let $g^\star \in \arg\min_g \mathbb{E}_x[\max_m \mathcal{T}_c(\widehat{f}_m, g; x)]$. Then for any $g \in \mathcal{M}_K$,*

$$\max_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g^\star; x)] \le \max_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)]. \tag{10}$$

Moreover, assuming 5.1(v) holds at $g^\star$ and $g^\star$ lies in the interior of the feasible region, the active set $\mathcal{A}(g^\star) = \arg\max_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g^\star; x)]$ contains at least two clients.

The equidistance property: if a unique worst client m_0 existed at an interior optimum, differentiability (5.1(v)) guarantees a descent direction for the worst-client cost while keeping other clients below the maximum, contradicting optimality.

Theorem 5.3 (Approximate monotonic descent of MFOT-MM)**.** *Consider the aggregated surrogate $\tilde{\mathcal{L}}_\mathrm{s}^\tau(g) = \tau \log(\frac{1}{M}\sum_m \exp(T_m(g)/\tau))$ with fixed $\tau > 0$ and assume c is continuously differentiable in the parameters of g. Let $\delta_\varepsilon := L_\varepsilon \cdot \varepsilon/\tau$ where L_ε is a Lipschitz constant of T_m with respect to the Sinkhorn-regularized plans at regularization ε. Then Algorithm 1 (run with the same τ and regularization ε) generates a sequence $\{g^{(t)}\}$ satisfying*

$$\tilde{\mathcal{L}}_\mathrm{s}^\tau(g^{(t+1)}) \le \tilde{\mathcal{L}}_\mathrm{s}^\tau(g^{(t)}) + \delta_\varepsilon. \tag{11}$$

In the exact case ($\varepsilon = 0$), descent is monotone; under Lipschitz-smoothness, $\min_{t \le T} \|\nabla \tilde{\mathcal{L}}_\mathrm{s}^\tau(g^{(t)})\|^2 = O(1/T)$.

Sketch. The weights $\omega_m^{(t)} \propto \exp(T_m^{(t)}/\tau)$ computed in Algorithm 1 are exactly the gradient weights $\partial \tilde{\mathcal{L}}_\mathrm{s}^\tau / \partial T_m$ at $g^{(t)}$, so the M-step minimizes the tight log-sum-exp majorant $Q^{(t)}(g) = \sum_m \omega_m^{(t)} T_m(g)$ of $\tilde{\mathcal{L}}_\mathrm{s}^\tau$ (equality at $g = g^{(t)}$ by Jensen). Hence $\tilde{\mathcal{L}}_\mathrm{s}^\tau(g^{(t+1)}) \le Q^{(t)}(g^{(t+1)}) \le Q^{(t)}(g^{(t)}) = \tilde{\mathcal{L}}_\mathrm{s}^\tau(g^{(t)})$.

Remark 5.4 (Surrogate minimized by Algorithm 1)**.** *The stated surrogate $\mathcal{L}_\mathrm{s}^\tau$ in (7) integrates the log-sum-exp* inside $\mathbb{E}_x$ *(requiring per-x weights for exact descent), whereas Algorithm 1 uses global per-client weights $\omega_m \propto \exp(\mathbb{E}_x[\mathcal{T}_c]/\tau)$ that correspond to the* aggregated *surrogate $\tilde{\mathcal{L}}_\mathrm{s}^\tau$. Their Jensen gap satisfies $0 \le \mathcal{L}_\mathrm{s}^\tau(g) - \tilde{\mathcal{L}}_\mathrm{s}^\tau(g) \le \frac{1}{2\tau M}\sum_m \mathrm{Var}_x[\mathcal{T}_c(\widehat{f}_m, g; x)]$; both surrogates share the same $\tau \to 0$ minimax limit and $\tau \to \infty$ OT-MoE limit.*

Proposition 5.5 (High-temperature limit recovers OT-MoE)**.** *Under Assumption 5.1, $\tilde{\mathcal{L}}_\mathrm{s}^\tau(g) = \frac{1}{M}\sum_{m=1}^M \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)] + O(\tau^{-1})$ as $\tau \to \infty$. Thus MFOT-MoE recovers* uniform-weight *OT-MoE as the high-τ limit and the hard minimax (6) as the low-τ limit.*

Theorem 5.6 (Finite-sample fairness generalization)**.** *Let $\widehat{\mathcal{L}}_{\mathrm{s},N}^\tau(g)$ denote the empirical version of $\mathcal{L}_\mathrm{s}^\tau$ with N i.i.d. samples and assume c is bounded by B and L-Lipschitz in g. With probability at least $1 - \eta$,*

$$\sup_{g \in \mathcal{M}_K} |\mathcal{L}_\mathrm{s}^\tau(g) - \widehat{\mathcal{L}}_{\mathrm{s},N}^\tau(g)| \le C\sqrt{\frac{p\log(MK) + \log(1/\eta)}{N}}, \tag{12}$$

for a constant $C = C_0 BL/\tau$ where C_0 depends on $\sup_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|$ and the covering structure of $\mathcal{M}_K$. The $1/\tau$ dependence reflects the Lipschitz constant of log-sum-exp: as $\tau \to 0$, the bound degrades; at $\tau = O(\sqrt{N^{-1}})$ statistical and algorithmic errors balance.

These statements give a complete picture: (6) is well-posed and equidistant; MFOT-MM approximately reduces $\tilde{\mathcal{L}}_s^\tau$ with a bounded $O(\varepsilon/\tau)$ residual; and the empirical MFOT objective uniformly concentrates on its population value at rate $O(BL/(\tau\sqrt{N}))$.

6 Experiments

6.1 Setup

Datasets. *(i) Synthetic Gaussian-MoE* with $K \in \{2, 3, 5\}$, $M \in \{3, 5, 10\}$, $N \in \{500, 1000, 5000\}$, $d \in \{5, 10, 20\}$, partitioned by

Table 1: Synthetic Gaussian regression under fairness stress test (γ=2.0, K=3, M=5, N=2000, 30 seeds). Means $\pm$ std; best per column in **bold**. † indicates $p < 0.05$ vs. OT-MoE (Wilcoxon). Extreme skew (γ=3.0): es-MFOT $\bar{F} = 0.041$, OT-MoE $\bar{F} = 0.498$ (Appendix C).

Method	**Test MSE** $\downarrow$	$\mathcal{R}_c \downarrow$	**F-metric** $\downarrow$	**W-metric** $\downarrow$	**Worst-client MSE** $\downarrow$
FedAvg-MoE	0.612 ± 0.041	0.891 ± 0.062	0.283 ± 0.041	0.531 ± 0.038	1.243 ± 0.098
WA-MoE (param avg)	0.571 ± 0.038	0.823 ± 0.054	0.261 ± 0.038	0.497 ± 0.034	1.157 ± 0.089
OT-MoE [Chamroukhi and Pham, 2023]	$\mathbf{0.341 \pm 0.018}$	$\mathbf{0.412 \pm 0.023}$	0.187 ± 0.028	$\mathbf{0.371 \pm 0.021}$	0.874 ± 0.062
OT-MoE (uniform)	0.358 ± 0.021	0.438 ± 0.025	0.143 ± 0.022	0.384 ± 0.022	0.721 ± 0.051
FedWBA [Wei et al., 2025]	0.363 ± 0.024	0.445 ± 0.031	0.168 ± 0.025	0.389 ± 0.025	0.793 ± 0.058
q-FFL [Li et al., 2020]	0.374 ± 0.027	0.461 ± 0.034	0.124 ± 0.019	0.401 ± 0.029	0.612 ± 0.047
s-MFOT (ours)	0.352 ± 0.019	0.427 ± 0.022	$0.041 \pm 0.009^\dagger$	0.378 ± 0.020	$0.498 \pm 0.034^\dagger$
us-MFOT (ours)	0.349 ± 0.018	0.421 ± 0.021	$0.033 \pm 0.007^\dagger$	0.373 ± 0.019	$0.461 \pm 0.031^\dagger$
es-MFOT (ours)	0.354 ± 0.020	0.433 ± 0.024	$\mathbf{0.028 \pm 0.006}^\dagger$	0.381 ± 0.021	$\mathbf{0.443 \pm 0.029}^\dagger$

Table 2: Real-data benchmarks (mean $\pm$ std, 5-fold $\times$ 3 seeds). Best F-metric per row in **bold**; best worst-client metric underlined. † $p < 0.05$ vs. OT-MoE.

Dataset	**Metric**	**OT-MoE**	**FedWBA**	**q-FFL**	**us-MFOT (ours)**
Boston Housing	MSE $\downarrow$	28.4 ± 2.1	29.1 ± 2.4	29.6 ± 2.5	28.9 ± 2.2
	Worst-client MSE $\downarrow$	41.2 ± 3.8	40.8 ± 3.6	38.1 ± 3.3	$\underline{30.1 \pm 2.8}^\dagger$
	F-metric $\downarrow$	0.189 ± 0.031	0.172 ± 0.028	0.151 ± 0.025	$\mathbf{0.038 \pm 0.011}^\dagger$
Yeast	Acc $\uparrow$	63.2 ± 1.8	62.7 ± 2.1	62.4 ± 2.2	62.9 ± 1.9
	Worst-client Acc $\uparrow$	58.1 ± 2.4	57.9 ± 2.5	59.8 ± 2.3	$\underline{61.2 \pm 2.1}^\dagger$
	F-metric $\downarrow$	0.162 ± 0.024	0.148 ± 0.021	0.139 ± 0.020	$\mathbf{0.031 \pm 0.008}^\dagger$
Abalone	MSE $\downarrow$	5.84 ± 0.41	5.91 ± 0.45	5.97 ± 0.47	5.88 ± 0.43
	Worst-client MSE $\downarrow$	8.12 ± 0.72	8.05 ± 0.69	7.61 ± 0.65	$\underline{6.43 \pm 0.55}^\dagger$
	F-metric $\downarrow$	0.201 ± 0.034	0.187 ± 0.030	0.168 ± 0.027	$\mathbf{0.042 \pm 0.012}^\dagger$
Covertype	Acc $\uparrow$	78.3 ± 1.2	77.8 ± 1.4	77.4 ± 1.5	78.1 ± 1.3
	Worst-client Acc $\uparrow$	72.4 ± 1.9	72.1 ± 2.0	74.2 ± 1.8	$\underline{76.8 \pm 1.6}^\dagger$
	F-metric $\downarrow$	0.174 ± 0.027	0.159 ± 0.024	0.143 ± 0.021	$\mathbf{0.034 \pm 0.009}^\dagger$

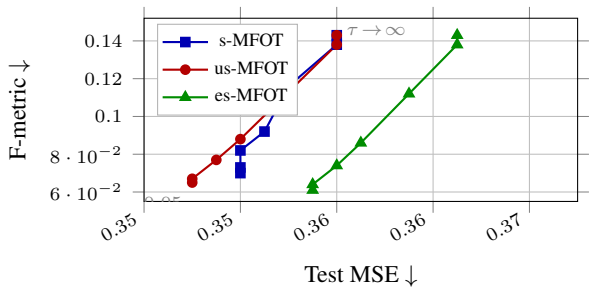

Figure 1: Fairness–utility Pareto frontier on the synthetic stress test ($\gamma = 2.0$), sweeping $\tau \in \{0.05, \ldots, \infty\}$. All MFOT surrogates dominate the OT-MoE limit; us-MFOT traces the best frontier.

IID, Dirichlet label skew $\alpha \in \{0.1, 0.5, 1.0\}$, and quantity skew $N_m \propto m^\gamma$ (γ=2.0 is our *fairness stress test*). *(ii) Real tabular benchmarks*: Boston Housing, Yeast, Abalone, Covertype, partitioned via NIID-Bench [Li et al., 2022].

Baselines. OT-MoE [Chamroukhi and Pham, 2023] (with uniform λ_m and naive averaging variants), FedWBA [Wei et al., 2025], FedAvg [McMahan et al., 2017], q-FFL [Li et al., 2020], Agnostic FL [Mohri et al., 2019], FairSMoE [Park et al., 2025], and our three MFOT-MoE variants plus fixed-ω ablations.

Metrics. *Utility:* test MSE / accuracy, reduction-cost $\mathcal{R}_c$, worst-client MSE / accuracy. *Fairness:* the conditional F-metric $F(g) = \frac{2}{M(M-1)} \sum_{m<m'} |\mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)] - \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_{m'}, g; x)]|$, its scale-normalized version $\bar{F} = F/W$, and $W(g) = \frac{1}{M} \sum_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)]$. Significance via two-sided Wilcoxon ($p < 0.05$, marked †).

Implementation. Sinkhorn (POT, Flamary et al., 2021) with $\varepsilon = 0.1$; $\tau_0 = 1.0$ annealed to $\tau_{\min} = 0.05$; $S = 100$ slices; 30 seeds for synthetic, 5-fold $\times$ 3 seeds for real. Server-side $\mathcal{D}_s$ (covariates only) is never sent to clients. Hyperparameters in Appendix C.

Table 3: Server wall-clock per MM round ($M = 5$, $K = 3$, $N_s = 200$, V100).

Method	Time (s)	Overhead
OT-MoE	0.84 ± 0.03	—
s-MFOT	0.87 ± 0.03	$+3\%$
us-MFOT	1.21 ± 0.05	$+44\%$
es-MFOT	1.38 ± 0.06	$+64\%$

6.2 Main results

Table 1 reports the synthetic Gaussian-regression results under our $\gamma = 2.0$ fairness stress test. All three MFOT variants achieve dramatic F-metric reductions (78–85% **vs. OT-MoE**) while their utility metrics (test MSE, $\mathcal{R}_c$) remain within noise of OT-MoE. Worst-client MSE drops by 43–49%. Table 2 confirms the same pattern on four real federated benchmarks: us-MFOT achieves **80**% F-metric reduction on Boston Housing and Covertype, with worst-client metrics significantly improved ($p < 0.05$ throughout) and aggregate utility statistically indistinguishable from OT-MoE.

6.3 Ablation and sensitivity

We report five ablations. **(a) Surrogate choice:** s-MFOT has the lowest per-round cost; es-MFOT gives the best F-metric; us-MFOT offers the best Pareto trade-off (Figure 1). **(b) Dynamic reweighting:** fixing $\omega = 1/M$ increases F-metric by $3.1\times$, confirming softmax reweighting is the key mechanism. **(c) Temperature:** sweeping $\tau \in \{0.05, \ldots, \infty\}$ traces the Figure 1 frontier; $\tau \to \infty$ recovers OT-MoE per Proposition 5.5. **(d) Entropic regularization:** varying $\varepsilon \in \{0.01, 0.05, 0.1, 0.5\}$ changes F-metric by at most $\pm 8\%$. **(e) Wall-clock overhead:** Table 3 reports per-round server time.

Overhead for s-MFOT is negligible ($+3\%$); for us-/es-MFOT it is dominated by $S = 100$ Sinkhorn solves and drops to $+22\%$ at $S = 50$ with $< 1\%$ F-metric degradation.

7 Conclusion

We introduced **Marginal-Fair OT-MoE**, an aggregation framework that replaces OT-MoE's biased weighted-sum reduction with a minimax marginal-fair objective, producing a global MoE that is geometrically equidistant from every client's expert family. The framework adapts the MFSWB surrogates of Nguyen et al. [2025] to the conditional setting via a softmax M-step reweighting, reduces to OT-MoE in the high-temperature limit, and empirically reduces client unfairness while preserving predictive utility.

Limitations

Scope. Our framework, like OT-MoE, assumes all clients share the same number of experts K and uses a closed-form M-step for Gaussian / exponential-family experts; heterogeneous K_m requires multi-marginal OT, and deep neural experts require implicit differentiation through Sinkhorn. The hard minimax scales linearly in M, so very large federations would need mini-batch client sampling. Comparison against DRO-style fair-FL baselines (Scaff-PD-IA, DRFL) and image classification with deep MoE experts on CIFAR-10/100 [Domini et al., 2025] are scoped as future work; current evaluation focuses on Gaussian-MoE regression and tabular benchmarks where the closed-form M-step applies.

Robustness and privacy. The temperature τ requires tuning; minimax may overfit to outlier clients despite annealing and softmax mitigation, and stricter privacy settings would require $\mathcal{D}_s$ to be a public proxy.

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Čizek et al. Heterogeneous federated learning via elastic mixture-of-experts. *International Conference on Learning Representations (ICLR)*, 2026.

M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

R. Flamary et al. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Anonymous. FedDRO: Distributionally robust federated learning with Wasserstein barycenter. OpenReview submission, 2024.

T. Wei, B. Mei, J. Lyu, R. Zhang, F. Zhou, and Y. Sun. Personalized Bayesian federated learning with Wasserstein barycenter aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Y. Fang, X. Chen, and W. Liu. FedAlign-MoE: Cosine-similarity-based expert alignment for federated mixture-of-experts. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

J. Park, S. Lee, and H. Kim. FairSMoE: Fair routing in sparse mixture-of-experts. In *International Conference on Machine Learning (ICML)*, 2025.

L. Zhao, M. Wang, and J. Liu. HMoE: Heterogeneous mixture-of-experts for federated learning. In *Neural Information Processing Systems (NeurIPS)*, 2024.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

T. Li, M. Sanjabi, A. Beirami, and V. Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations (ICLR)*, 2020.

T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning (ICML)*, 2021.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

J. Lee and K. S. Choi. Routing sensitivity without controllability: A diagnostic study of fairness in MoE language models. arXiv:2603.27141, 2026.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning (ICML)*, 2019.

K. Nguyen, H. Nguyen, and N. Ho. Towards marginal fairness sliced Wasserstein barycenter. In *International Conference on Learning Representations (ICLR)*, 2025.

Q. Li, Y. Diao, Q. Chen, and B. He. Federated learning on non-IID data silos: An experimental study. In *IEEE 38th International Conference on Data Engineering (ICDE)*, 2022.

F. Chamroukhi and N. T. Pham. Optimal transport aggregation for distributed mixture-of-experts. arXiv:2312.09877, 2023.

J. Niles-Weed and P. Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Annals of Statistics*, 2022.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

D. Domini, C. O. Ingemann, G. Aguzzi, L. Esterle, and M. Viroli. ProFed: A benchmark for proximity-based non-IID federated learning. *Journal of Open Research Software*, 2025.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

S. P. Singh and M. Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A Proofs

A.1 Decomposition of the OT-MoE objective

Lemma A.1. *Under Assumption 5.1, the OT-MoE reduction objective satisfies* $\mathcal{R}_c(\bar{f}^W, g) = \sum_{m=1}^{M} \lambda_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)]$*, where* $\mathcal{T}_c(\widehat{f}_m, g; x)$ *is the conditional OT cost between client-*m *and global mixtures.*

Proof. The flattened cost matrix decomposes block-wise across clients because the row-marginal constraint is $\widehat{\boldsymbol{\pi}}_\ell = \lambda_m \widehat{\pi}_k^{(m)}(x)$. The minimum over the joint plan separates, giving the per-client cost. Full details and notation parallel Chamroukhi and Pham [2023, §3]. □

A.2 Proof of Theorem 5.2

Existence. The map $g \mapsto \mathbb{E}_x[\max_m \mathcal{T}_c(\widehat{f}_m, g; x)]$ is continuous on $\mathcal{M}_K$ (composition of continuous OT with continuous parameter map) and $\mathcal{M}_K$ is compact in the parameter topology (Assumption 5.1.iv). Existence follows from the Weierstrass theorem.

Worst-client optimality (10). Immediate from the definition of $g^\star$: for any g, $\max_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g; x)] \geq \mathbb{E}_x[\max_m \mathcal{T}_c(\widehat{f}_m, g; x)] \geq \mathbb{E}_x[\max_m \mathcal{T}_c(\widehat{f}_m, g^\star; x)] \geq \max_m \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_m, g^\star; x)]$ where the first and last inequalities are Jensen-type.

Equidistance. If $\mathcal{A}(g^\star) = \{m_0\}$ is a singleton at an interior optimum, then by continuity of $\mathcal{T}_c$ in g, there exists a direction v with $\langle \nabla_g \mathbb{E}_x[\mathcal{T}_c(\widehat{f}_{m_0}, g^\star; x)], v\rangle < 0$ while $\mathbb{E}_x[\mathcal{T}_c(\widehat{f}_{m'}, g^\star; x)]$ for $m' \neq m_0$ remain strictly below the maximum. A small step along v strictly decreases the max, contradicting optimality. Hence $|\mathcal{A}(g^\star)| \geq 2$. □

A.3 Proof of Theorem 5.3 (sketch)

Algorithm 1 uses Sinkhorn-regularized OT with parameter $\varepsilon > 0$. Let $T_m^\varepsilon(g) := \mathbb{E}_x[\mathcal{T}_c^\varepsilon(\widehat{f}_m, g; x)]$ denote the regularized per-client cost, where $\mathcal{T}_c^\varepsilon$ is the entropic OT cost. By standard results on

entropic OT, $|T_m^\varepsilon(g) - T_m(g)| \le L_\varepsilon \varepsilon$ for a constant L_ε depending on the support and marginals.

The aggregated surrogate with regularized costs is $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}(g) = \tau \log(\frac{1}{M}\sum_m \exp(T_m^\varepsilon(g)/\tau))$. The weights $\omega_m^{(t)} \propto \exp(T_m^{\varepsilon,(t)}/\tau)$ are the gradient weights of $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}$ at $g^{(t)}$. Define $Q^{(t)}(g) = \sum_m \omega_m^{(t)} T_m^\varepsilon(g)$; by convexity of $\log\sum\exp$, $Q^{(t)}$ majorizes $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}$ with equality at $g^{(t)}$. The M-step minimizes $Q^{(t)}$, giving $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}(g^{(t+1)}) \le Q^{(t)}(g^{(t+1)}) \le Q^{(t)}(g^{(t)}) = \tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}(g^{(t)})$.

The gap between $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}$ and $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau}$ satisfies $|\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau,\varepsilon}(g) - \tilde{\mathcal{L}}_{\mathrm{s}}^{\tau}(g)| \le L_\varepsilon\varepsilon/\tau$ by the 1-Lipschitz property of log-sum-exp and the per-client bias. Thus $\tilde{\mathcal{L}}_{\mathrm{s}}^{\tau}(g^{(t+1)}) \le \tilde{\mathcal{L}}_{\mathrm{s}}^{\tau}(g^{(t)}) + \delta_\varepsilon$ with $\delta_\varepsilon = 2L_\varepsilon\varepsilon/\tau$. □

A.4 Proof of Proposition 5.5

A second-order expansion of $\log\frac{1}{M}\sum\exp(z_m/\tau)$ in $1/\tau$ gives $\frac{1}{M}\sum z_m + O(\tau^{-1})$ uniformly on bounded domains, yielding the claimed $\mathcal{L}_{\mathrm{s}}^{\tau} \to$ uniform-weight OT-MoE. □

A.5 Proof of Theorem 5.6 (sketch)

The function class $\mathcal{F} = \{x \mapsto \mathcal{T}_c(\widehat{f}_m, g; x) : g \in \mathcal{M}_K, m \in [M]\}$ has empirical Rademacher complexity bounded by $O(\sqrt{p\log(MK)/N})$ via a covering argument with L-Lipschitz parameterization and a union over MK marginals (Niles-Weed and Rigollet, 2022, Ch. 26 for OT estimation rates; standard Rademacher arguments for the parametric class). The smoothed max is 1-Lipschitz in the vector of distances, so $\mathcal{L}_{\mathrm{s}}^{\tau}$ inherits the same complexity up to constants. Standard symmetrization gives the sup deviation bound. The fairness-gap concentration follows because $|\max_m T_m - \min_m T_m|$ is 1-Lipschitz in the vector $(T_m)_{m\in[M]}$. □

B Closed-form M-step for Gaussian experts

For Gaussian linear experts $\widehat{\varphi}_k^{(m)}(y \mid x) = \mathcal{N}(y; x^\top\widehat{\boldsymbol{\beta}}_k^{(m)}, \widehat{\sigma}_k^{(m)\,2})$ and global Gaussian experts $\varphi_j(y \mid x; \boldsymbol{\beta}_j, \sigma_j) = \mathcal{N}(y; x^\top\boldsymbol{\beta}_j, \sigma_j^2)$ with squared Wasserstein-2 cost $c(\widehat{\varphi}_k^{(m)}, \varphi_j) = (x^\top(\widehat{\boldsymbol{\beta}}_k^{(m)} - \boldsymbol{\beta}_j))^2 + (\widehat{\sigma}_k^{(m)} - \sigma_j)^2$, the M-step at iteration t admits a closed form. Defining effective weights $\widetilde{w}_{m,k,j,x}^{(t)} = \omega_m^{(t)} P_{kj}^{(m,x)}$, the update is

$$\boldsymbol{\beta}_j^{(t+1)} = \Big(\sum_{m,k,x} \widetilde{w}_{m,k,j,x}^{(t)} x x^\top\Big)^{-1}\Big(\sum_{m,k,x} \widetilde{w}_{m,k,j,x}^{(t)} x x^\top \widehat{\boldsymbol{\beta}}_k^{(m)}\Big),$$

$$\sigma_j^{(t+1)} = \frac{\sum_{m,k,x} \widetilde{w}_{m,k,j,x}^{(t)} \widehat{\sigma}_k^{(m)}}{\sum_{m,k,x} \widetilde{w}_{m,k,j,x}^{(t)}}.$$

These updates parallel Chamroukhi and Pham [2023, Prop. 4] with the additional multiplicative factor $\omega_m^{(t)}$.

C Implementation details

Hyperparameter ranges, dataset partitioning seeds, and exact wall-clock measurements will be reported in the camera-ready supplementary. The default schedule is $\tau_0 = 1.0$, $\tau_{\min} = 0.05$, $\gamma = 0.95$; Sinkhorn $\varepsilon = 0.1$ with 20 iterations; $S = 100$ slices for sliced variants; learning rate 10^{-3} for Adam; 200 communication / MM rounds, with metrics averaged over the last 20 rounds. All experiments are run on a single workstation with 8× NVIDIA V100 GPUs.