Federated Distributionally Robust Neural Combinatorial Optimization with Convergence Guarantees

Ryunosuke Hamada
DENSO CORPORATION
ryunosuke.hamada.j3h@jp.denso.com

Daisuke Hatano
DENSO CORPORATION
daisuke.hatano.j6c@jp.denso.com

Yuichi Takano
University of Tsukuba
ytakano@sk.tsukuba.ac.jp

Abstract

Neural combinatorial optimization (NCO) solvers generalize poorly under distribution shift, yet robustness across heterogeneous problem distributions is essential for real-world deployment where no single client should be left behind. Group distributionally robust optimization (Group DRO) offers a principled remedy, but requires pooling data that often cannot be shared due to confidentiality or regulatory constraints. Federated learning removes data-sharing requirements; however, existing federated DRO methods cannot accommodate the policy-gradient estimators used in modern NCO. Specifically, prior analyses rely on Lipschitz losses and bounded gradient variance. The REINFORCE policy-gradient estimator multiplies sampled costs by score functions, so bounded costs alone do not guarantee finite gradient variance. We address this gap with explicit smoothness, bounded-gradient, unbiasedness, and variance assumptions for the REINFORCE estimator. Under these assumptions, we establish the first federated Group DRO framework with an inverse-square-root convergence rate in the number of communication rounds. Our analysis applies to any REINFORCE-trained model under these assumptions, including NCO solvers as a special case. Experiments on capacitated vehicle routing demonstrate that an independently trained minority-client model has 26% higher overall validation cost than federated models, illustrating the value of federation for cross-client generalization. We show that unregularized Group DRO can suffer weight collapse, while KL regularization prevents cumulative collapse of the weights; weights may nonetheless concentrate when loss differences are large relative to the temperature parameter. The KL-regularized Group DRO objective yields a formal worst-group upper bound, and results in the tested settings are comparable to standard federated averaging.

1 Introduction

Neural combinatorial optimization (NCO) trains neural networks end-to-end to solve combinatorial optimization problems such as vehicle routing, producing high-quality solutions orders of magnitude faster than classical solvers [Kool et al., 2019, Kwon et al., 2020]. A central limitation is that these models are optimized for a single problem distribution: under distribution shift—different cities, customer densities, or demand patterns—performance degrades substantially [Joshi et al., 2022, Bi et al., 2022]. In logistics, independent operators naturally hold heterogeneous private distributions that cannot be pooled due to confidentiality or regulatory constraints. Independent local training on

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

such small private datasets fails to generalize across operators, motivating a federated approach with formal robustness guarantees.

We formalize this as a federated group distributionally robust optimization (Group DRO) problem for policy-gradient NCO: given multiple clients, each holding a private instance distribution, find a shared policy that minimizes the worst-group expected cost, communicating only model parameters and tolerating partial client participation (clients may be offline or bandwidth-constrained in any given round). Empirical risk minimization (ERM)—minimizing the uniform average across clients—provides no guarantee on worst-group cost [Sagawa et al., 2020]. KL-regularized Group DRO (hereafter KL-DRO) addresses this by minimizing a smooth surrogate that upper-bounds worst-group cost and interpolates between near-worst-case robustness and ERM via a temperature parameter (Section 3). We seek algorithms that converge at the standard inverse-square-root rate in the number of communication rounds under this federated setting.

The following two challenges make this difficult in practice. **Challenge 1: REINFORCE variance depends jointly on costs and score functions.** Policy-gradient estimators multiply the sampled cost by the score function [Williams, 1992], so gradient variance scales with squared costs and score-function norms [Greensmith et al., 2004] rather than being bounded by a Lipschitz constant as in supervised learning [Bottou et al., 2018]. This breaks the variance assumptions on which existing federated DRO convergence proofs rely. **Challenge 2: Unregularized group weights collapse.** The exponentiated-gradient update of standard GroupDRO [Sagawa et al., 2020] concentrates weight exponentially fast onto the single highest-loss client. Although weight concentration is a known property of multiplicative-weight methods in the bandit and online learning literature [Arora et al., 2012, Lattimore and Szepesvári, 2020], its severity in neural combinatorial optimization—and the resulting worst-group degradation—has not been formally characterized or empirically isolated from federation effects.

These challenges expose a clear gap in the literature. In centralized NCO, Jiang et al. [2022] applied Group DRO to routing but used heuristic weight updates without convergence guarantees, and their unregularized formulation is susceptible to Challenge 2. In federated DRO, Guo and Yang [2024] established an inverse-square-root convergence rate for KL-regularized objectives (FGDRO-KL) under Lipschitz-loss and bounded-variance assumptions, with full client participation required. Their analysis does not address the REINFORCE variance issue in Challenge 1. No prior work provides a convergence-guaranteed federated DRO framework for policy-gradient solvers.

Our key insight is that *bounded costs control the cost factor in the REINFORCE gradient estimator*. In the CVRP setting studied here, route costs are bounded; more generally, finite gradient variance requires a separate bound on the score-function moment. Bounded costs alone therefore do not guarantee bounded gradient variance. Under the stated smoothness, bounded-gradient, unbiasedness, and variance assumptions, we establish smoothness of the DRO surrogate and convergence for REINFORCE-trained policies. KL regularization gives every client a strictly positive weight in exact arithmetic when its reference weight is positive and the temperature is finite, although a weight may still be numerically negligible or highly concentrated. A server-computed normalization constant supports aggregation under partial participation; the EMA convergence analysis, however, assumes full participation. Building on these ideas, we propose two algorithms based on federated averaging (FedAvg) [McMahan et al., 2017]—an idealized variant (Algorithm 2) and a communication-efficient exponential moving average (EMA) variant (Algorithm 1). Our main contributions can be summarized as follows.

1. **Federated DRO framework for policy-gradient solvers** (Algorithms 1, 2). We propose the first federated Group DRO framework applicable to REINFORCE-trained models, comprising an idealized variant and a communication-efficient EMA variant with server-managed normalization for partial client participation.
2. **Convergence guarantees** (Theorems 1, 2; Corollary A.1). The idealized algorithm has an inverse-square-root stationarity bound under fixed-size uniform partial participation. For the EMA algorithm, we prove a finite-time descent bound and, through coupled tracking, an inverse-square-root stationarity rate with full participation, one local step per communication round, and horizon-dependent parameters fixed within each run.
3. **Weight-collapse analysis** (Proposition 1). We prove that unregularized GroupDRO weights collapse exponentially onto a single client under a persistent positive loss gap. In experiments, aggressive GroupDRO step sizes produce 5–26% worst-group degradation across the 12 federated

conditions. A centralized baseline also exhibits collapse and 3.3% degradation, indicating that the phenomenon is not limited to federated training.

4. **Empirical insights** (Section 6). Federated training prevents the 26% overall degradation of independent local training for minority clients. KL-DRO has comparable worst-group results to ERM in the tested settings, while the objective supplies a formal worst-group upper bound. We observe near-orthogonality of inter-group REINFORCE gradients in the capacitated vehicle routing problem (CVRP), a property that persists even under 100-fold variance reduction.

For reference, notation used throughout the paper is summarized in Appendix A.2.

2 Related work

Neural combinatorial optimization. Learning-based approaches to combinatorial optimization have progressed from pointer networks [Vinyals et al., 2015] and sequence-to-sequence models [Nazari et al., 2018] to REINFORCE-trained attention-based architectures [Vaswani et al., 2017, Kool et al., 2019] and their extensions such as POMO [Kwon et al., 2020]; see Angioni et al. [2025] for a recent tutorial and Bengio et al. [2021] for a broader survey. These methods achieve near-optimal solutions on uniform random instances but exhibit limited robustness under distribution shift [Joshi et al., 2022, Bi et al., 2022]. Recent efforts address this fragility through diverse strategies including efficient active search [Hottung et al., 2022], learning generalizable representations across problem sizes [Drakulic et al., 2023], and mixture-of-experts architectures [Zhou et al., 2024]. Liu et al. [2024] introduced a federated graph-neural-network framework for the traveling salesperson problem, and Meng et al. [2026] concurrently proposed federated learning to share knowledge across vehicle-routing problem types; these works focus on federated training architectures and knowledge transfer, whereas our work targets the orthogonal problem of distributional robustness with convergence guarantees under policy-gradient estimators.

Federated learning (FL). FedAvg [McMahan et al., 2017] reduces communication via multiple local stochastic gradient descent (SGD) steps, and its variants (SCAFFOLD [Karimireddy et al., 2020], FedProx [Li et al., 2020a], FedAdam [Reddi et al., 2021]) further improve convergence under data heterogeneity; see Kairouz et al. [2021] for a survey. The convergence of local SGD under data heterogeneity has been studied extensively [Stich, 2019, Li et al., 2020b, Karimireddy et al., 2020, Khaled et al., 2020], with client drift as a central challenge. Personalized FL approaches such as Ditto [Li et al., 2021] address heterogeneity via per-client models; in contrast, our DRO objective seeks a single robust global policy. Khodadadian et al. [2022] extended federated convergence theory to value-based reinforcement learning (RL) with linear speedup under Markovian sampling, but did not address policy-gradient estimators or DRO objectives. Our analysis extends this line of work to the DRO setting with REINFORCE gradient estimators.

Group DRO and federated DRO. Group DRO [Sagawa et al., 2020, Namkoong and Duchi, 2016] optimizes worst-case performance over predefined groups; we refer to Rahimian and Mehrotra [2022] and Kuhn et al. [2025] for broader DRO overviews, and to Hashimoto et al. [2018] for connections to fairness. Jiang et al. [2022] applied Group DRO to routing in a centralized setting with a convolutional neural network module for learning distribution-specific node patterns; in contrast, we operate in a federated setting with a distinct policy-gradient regime and provide non-asymptotic convergence rates. KL-regularized Group DRO yields a smooth surrogate amenable to gradient-based optimization; stochastic approximation approaches to Group DRO are studied by Zhang et al. [2023], while related variance-based regularization is studied by Duchi and Namkoong [2019]. In the federated setting, Mohri et al. [2019] proposed agnostic FL for worst-case client performance, and Deng et al. [2020] proposed federated DRO (DRFA) with adaptive client sampling and periodic model averaging for a min–max objective over client losses, using $\mathcal{O}(T_{\mathrm{iter}}^{3/4})$ communication rounds over T_{iter} optimization iterations. Guo and Yang [2024] analyzed FGDRO-KL, establishing a communication complexity of $\mathcal{O}(1/\epsilon^3)$ to reach a desired precision level ϵ, under full client participation with supervised SGD estimators. We adapt FGDRO-KL to the policy-gradient regime with three key differences: (i) explicit REINFORCE estimator and moment conditions, with bounded costs for EMA tracking (Remark A.5), (ii) server-managed normalization with separate analyses of idealized partial participation and full-participation EMA tracking (Remark B.4), and (iii) a theoretical analysis of weight collapse in unregularized GroupDRO (Proposition 1). Related risk-sensitive objectives include CVaR-based

policy gradients [Tamar et al., 2015], but these have not been studied in federated or group-DRO settings for combinatorial optimization.

3 Problem formulation

Our goal is to train a single neural policy that performs well for every client, even when their problem distributions differ and data cannot be shared. Standard federated averaging minimizes the average cost but offers no guarantee on the worst-performing client. Directly minimizing the worst-group cost is challenging for gradient-based methods because the pointwise maximum is non-smooth. KL-regularized Group DRO resolves this by replacing the non-smooth maximum with a smooth surrogate that upper-bounds the worst-group cost after an additive correction; this yields closed-form softmax weights, enabling standard SGD convergence analysis while still providing a formal robustness certificate on the worst group. Below we formalize these components.

3.1 Federated neural combinatorial optimization

We consider a general combinatorial optimization (CO) setting with G clients, indexed by $g \in \mathcal{G} = \{1, 2, \ldots, G\}$, each holding a private instance distribution $\mathcal{D}_g$. A problem instance x is drawn from a client distribution, and a feasible solution τ incurs a cost $\ell(\tau) \in \mathbb{R}_+$. We parameterize a stochastic policy $\pi_\theta(\tau \mid x)$ by a neural network with parameters $\theta \in \mathbb{R}^D$ that autoregressively constructs a solution τ given instance x. The expected cost under distribution $\mathcal{D}_g$ for client g is

$$L_g(\theta) = \mathbb{E}_{x \sim \mathcal{D}_g} \, \mathbb{E}_{\tau \sim \pi_\theta(\cdot|x)} \big[\ell(\tau)\big]. \tag{1}$$

Throughout the experiments we focus on the capacitated vehicle routing problem (CVRP), which seeks minimum-cost routes for capacity-constrained vehicles serving customers from a central depot. An instance x specifies customer locations, demands, and vehicle capacity; a feasible solution τ is a set of routes respecting capacity constraints, and $\ell(\tau)$ is the total travel distance.

We require the following two assumptions on the problem and the expected loss landscape.

Assumption 1 (Bounded cost). $|\ell(\tau)| \le B_L$ a.s. for all $g \in \mathcal{G}$, $\theta \in \mathbb{R}^D$, $x \sim \mathcal{D}_g$, and $\tau \sim \pi_\theta(\cdot \mid x)$.

Assumption 2 (Smoothness of client losses). Each $L_g(\theta)$ is M_L-smooth: $\|\nabla L_g(\theta_1) - \nabla L_g(\theta_2)\| \le M_L \|\theta_1 - \theta_2\|$ for all $g \in \mathcal{G}$ and $\theta_1, \theta_2 \in \mathbb{R}^D$.

For CVRP, customer and depot locations lie in a bounded region and each instance has finitely many customers, so every feasible route has bounded cost. More broadly, Assumption 1 applies whenever the environment costs are bounded, as is standard in policy-gradient RL [Sutton and Barto, 2018]. Assumption 2 is a regularity condition on the expected loss landscape; we impose it rather than derive it from a bounded-weights condition on the policy network. Together with the gradient and variance assumptions in Section 5, these conditions are the key departure from prior federated DRO analyses, which instead require the *stochastic* loss to be Lipschitz. The algorithmic framework and convergence analysis (Sections 4–5) apply to any policy-gradient RL problem satisfying the stated assumptions, covering CO tasks such as TSP, job scheduling, and bin packing.

The standard federated ERM objective minimizes the average loss across all clients:

$$\min_{\theta \in \mathbb{R}^D} \; \frac{1}{G} \sum_{g=1}^{G} L_g(\theta). \tag{2}$$

ERM treats all groups equally, which can lead to poor worst-case performance when client distributions are heterogeneous [Sagawa et al., 2020].

3.2 KL-regularized group DRO (KL-DRO)

To promote robustness across groups, we consider the KL-DRO objective with reference distribution $p \in \Delta_G$:

$$\min_{\theta \in \mathbb{R}^D} \; \max_{q \in \Delta_G} \; \sum_{g=1}^{G} q_g L_g(\theta) - \lambda \sum_{g=1}^{G} q_g \log \frac{q_g}{p_g}, \tag{3}$$

where $\lambda > 0$ is a temperature parameter and $\Delta_G = \{q \in \mathbb{R}_+^G : \sum_g q_g = 1\}$ is the probability simplex. For positive reference weights, the KL penalty gives strictly positive optimal weights at finite $\lambda > 0$. Large λ drives the weights toward the reference p (recovering ERM when p is uniform), while small λ approaches the worst-case objective $\min_\theta \max_g L_g(\theta)$.

Smooth DRO surrogate. Since the inner maximization is strictly concave in q, the optimal adversarial weights admit a closed-form softmax solution (derived in Appendix A.5 for completeness):

$$q_g^\star(\theta) = \xi_g(\theta) := \frac{p_g \exp(L_g(\theta)/\lambda)}{\sum_{h=1}^G p_h \exp(L_h(\theta)/\lambda)}. \tag{4}$$

Substituting back, the min-max problem (3) reduces to minimizing the smooth surrogate:

$$\phi(\theta) := \lambda \log\left(\sum_{g=1}^G p_g \exp\left(\frac{L_g(\theta)}{\lambda}\right)\right). \tag{5}$$

The gradient of ϕ is a convex combination of client gradients with softmax weights: $\nabla\phi(\theta) = \sum_{g=1}^G \xi_g(\theta)\, \nabla L_g(\theta)$. Moreover, the worst-group loss is bounded as $\max_g L_g(\theta) \le \phi(\theta) - \lambda \log(\min_g p_g)$ (Lemma A.1 in Appendix).

Weight collapse in naïve GroupDRO. The KL penalty in the KL-DRO problem (3) addresses weight collapse. Without it, the standard exponentiated-gradient update used by GroupDRO [Sagawa et al., 2020] concentrates weights exponentially onto a persistently highest-loss group with a positive loss gap—a well-known property of multiplicative-weight updates [Arora et al., 2012, Lattimore and Szepesvári, 2020], but one whose consequences for GroupDRO training have not been formally quantified:

Proposition 1 (GroupDRO weight collapse). Under the exponentiated-gradient update $q_g^{(t+1)} \propto q_g^{(t)} \exp(\eta_q L_g(\theta^{(t)}))$ with $q^{(0)} = p$, if the highest-loss group $g^\star$ is constant across rounds with minimum gap $\delta_{\min} := \min_{0 \le t < T}\big(L_{g^\star}(\theta^{(t)}) - \max_{g \ne g^\star} L_g(\theta^{(t)})\big) > 0$, then $q_{g^\star}^{(T)} \ge 1 - \frac{1-p_{g^\star}}{p_{g^\star}} \exp(-\eta_q \delta_{\min} T)$.

Under these conditions, the weight on $g^\star$ converges to 1 exponentially fast, effectively discarding all other groups. In contrast, the KL-regularized weights in Eq. (4) depend only on the *current* losses—not on accumulated past losses. Since costs are bounded by B_L (Assumption 1), every weight is bounded below by $\xi_g \ge p_{\min} \exp(-B_L/\lambda) > 0$ for positive $p_{\min}$ and finite $\lambda > 0$. This bound is uniform across rounds but can be numerically negligible, so it does not rule out effective concentration (Remark B.2). The additive-centering ablation illustrates such concentration at $\lambda = 2$ on the C11 configuration (Appendix A.3). See Appendix B.1 for the proof.

4 Algorithm

We propose two federated algorithms for minimizing the KL-DRO surrogate in Eq. (5). The idealized variant assumes exact DRO weights (Algorithm 2 in Appendix A.1); the communication-efficient EMA variant estimates them from local minibatch costs (Algorithm 1). Their $\mathcal{O}(T^{-1/2})$ stationarity guarantees have different scopes: Theorem 1 (idealized) permits fixed-size uniform partial participation, whereas Corollary A.1 (EMA) requires full participation, $K = 1$, and horizon-dependent parameter choices. Both algorithms follow the FedAvg protocol: at each round t, a subset $S^{(t)} \subset \mathcal{G}$ of clients receive $\theta^{(t)}$, perform K local gradient steps (K is the number of local steps per round), and return updated models. The idealized analysis uses a fixed subset size m and uniform sampling without replacement (Assumption 6). The key modification over standard FedAvg is that each client's REINFORCE gradient is scaled by the softmax DRO weight $\xi_g(\theta)$ defined in Eq. (4), upweighting high-cost groups. Ideally, client g would use the exact weight to form the scaled gradient

$$\hat{d}_{g,k}^{(t)} = \xi_g(\theta_{g,k-1}^{(t)}) \cdot \frac{1}{B} \sum_{i=1}^B \ell(\tau_i)\, \nabla_\theta \log \pi_\theta(\tau_i \mid x_i), \tag{6}$$

Algorithm 1 FedDRO-KL-EMA for NCO

Require: Initial $\theta^{(0)}$, $v^{(0)}$; learning rate η; local steps K; EMA parameters $\beta_u, \beta_v, \beta_d$

for $t = 0, 1, \ldots, T-1$ **do**
  Broadcast $\theta^{(t)}$, $v^{(t)}$ to all clients $g \in S^{(t)}$
  for each client $g \in S^{(t)}$ **in parallel do**
    Initialize $\theta_{g,0}^{(t)} \leftarrow \theta^{(t)}$; load EMA states $u_{g,0}^{(t)}, \tilde{v}_{g,0}^{(t)}, \hat{d}_{g,0}^{(t)}$ from previous round[1]
    for $k = 1, \ldots, K$ **do**
      Sample minibatch; compute $\bar{\ell}_{g,k}^{(t)}$ and $\widehat{\nabla L_g}$ (REINFORCE)
      Update $u_{g,k}^{(t)}, \tilde{v}_{g,k}^{(t)}, \hat{d}_{g,k}^{(t)}$ via Eqs. (7)–(9)
      $\theta_{g,k}^{(t)} \leftarrow \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}$
    end for
    Send $\theta_{g,K}^{(t)}, \tilde{v}_{g,K}^{(t)}$ to server; retain $u_{g,K}^{(t)}, \tilde{v}_{g,K}^{(t)}, \hat{d}_{g,K}^{(t)}$ locally
  end for
  $\theta^{(t+1)} \leftarrow \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \theta_{g,K}^{(t)}$; $\quad v^{(t+1)} \leftarrow \frac{G}{|S^{(t)}|} \sum_{g \in S^{(t)}} \tilde{v}_{g,K}^{(t)}$
end for

where $x_i \sim \mathcal{D}_g$, $\tau_i \sim \pi_\theta(\cdot \mid x_i)$, and the model is updated as $\theta_{g,k}^{(t)} = \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}$. The policy and its score are evaluated at the current local iterate; the minibatch is drawn afresh conditional on the history before that step. The theoretical estimator is baseline-free and unclipped. Baseline subtraction is permitted under the conditional-independence and moment conditions in Remark A.4. This idealized procedure (Algorithm 2, Appendix A.1) serves as a theoretical reference but requires knowledge of all clients' true losses $L_g(\theta)$ for computing ξ_g.

In practice, each client observes only its own minibatch costs. Algorithm 1 replaces the exact weights with three exponential moving averages (EMAs) per client, following Guo and Yang [2024]: (i) a loss estimate u_g tracking the running mean cost, (ii) a local normalization $\tilde{v}_g$ tracking $p_g \exp(u_g/\lambda)$, and (iii) a gradient direction $\hat{d}_g$ accumulating the DRO-weighted REINFORCE signal. The server aggregates client models and local $\tilde{v}_g$ states into an approximation of the normalization constant. The protocol permits partial participation, but the EMA analysis here is limited to full participation and $K = 1$. The update rules with EMA parameters $\beta_u, \beta_v, \beta_d \in (0, 1]$ are:

$$u_{g,k}^{(t)} \leftarrow (1-\beta_u)\, u_{g,k-1}^{(t)} + \beta_u\, \bar{\ell}_{g,k}^{(t)}, \tag{7}$$

$$\tilde{v}_{g,k}^{(t)} \leftarrow (1-\beta_v)\, \tilde{v}_{g,k-1}^{(t)} + \beta_v\, p_g \exp(u_{g,k}^{(t)}/\lambda), \tag{8}$$

$$\hat{d}_{g,k}^{(t)} \leftarrow (1-\beta_d)\, \hat{d}_{g,k-1}^{(t)} + \beta_d\, \frac{p_g \exp(u_{g,k}^{(t)}/\lambda)}{v^{(t)}}\, \widehat{\nabla L_g}, \tag{9}$$

where $\bar{\ell}_{g,k}^{(t)}$ is the batch mean cost, $\widehat{\nabla L_g}$ is the REINFORCE gradient from the same minibatch, and $v^{(t)}$ is the server-aggregated normalization constant from the previous round. In the analysis, $\widehat{\nabla L_g}$ is the baseline-free, unclipped minibatch estimator of Appendix A.7. Its conditional unbiasedness does not imply unbiasedness of the product in Eq. (9), since the EMA weight depends on the current costs. The local model is updated as $\theta_{g,k}^{(t)} = \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}$. At each communication round, the server averages client models into $\theta^{(t+1)}$ and aggregates local weight estimates into $v^{(t+1)} = \frac{G}{|S^{(t)}|} \sum_{g \in S^{(t)}} \tilde{v}_{g,K}^{(t)}$; all three EMA states are retained locally across rounds. This normalization is not generally an unbiased estimate of $Z(\theta^{(t)})$: loss estimation, exponentiation, and state retention introduce tracking error.

Practical considerations. The experiments use divisive logits $z_g = (u_g/\bar{u} - 1)/\lambda$, where $\bar{u} > 0$ is the mean of the retained client loss estimates. At a given evaluation of these weights, this corresponds to an effective temperature $\lambda_{\text{eff}} = \lambda\bar{u}$, which changes during training. It is a heuristic change of weighting, not merely numerical stabilization of the fixed-λ objective analyzed below. In contrast,

[1] At t=0: $u_{g,0}^{(0)}$ is set to a minibatch estimate of $L_g(\theta^{(0)})$, $\tilde{v}_{g,0}^{(0)} = p_g e^{u_{g,0}^{(0)}/\lambda}$, $\hat{d}_{g,0}^{(0)} = \mathbf{0}$; for $t \geq 1$: carried over from $u_{g,K}^{(t-1)}, \tilde{v}_{g,K}^{(t-1)}, \hat{d}_{g,K}^{(t-1)}$. The analysis additionally requires an unbiased initial loss estimate with $|u_{g,0}^{(0)}| \leq B_L$ and mean-square error at most σ_ℓ^2, and $v^{(0)} = \sum_g \tilde{v}_{g,0}^{(0)}$ (Appendix A.13.1). See Algorithm 7 for expanded pseudocode.

additive centering $z'_g = (u_g - \max_h u_h)/\lambda$ preserves normalized exponential weights at fixed λ. We clip gradient norms to 1.0 and use $K = 1$ for the EMA experiments. Full details (centering logits, gradient clipping, momentum) appear in Appendix A.3. The additive-centering ablation is reported in Appendix A.3.

5 Convergence analysis

We now formalize the convergence properties previewed in Section 4. The main analytic challenge is that the softmax weighting in ϕ amplifies curvature by GH^2/λ, coupling smoothness to the robustness parameter. Smoothness of the expected losses and bounded true gradients control this curvature; bounded costs additionally control EMA loss and weight tracking. Both analyses use fresh REINFORCE estimators under the regularity and sampling conditions in Fact A.1 and Proposition A.1. The assumptions below quantify heterogeneity and noise; they are needed for the rate bound but do not affect algorithm operation. All proofs are deferred to the Appendix.

5.1 Smoothness of the DRO surrogate

Assumption 3 (Bounded gradients). $\|\nabla L_g(\theta)\| \le H$ for all $g \in \mathcal{G}$ and $\theta \in \mathbb{R}^D$.

This is an independent assumption on true gradients, not a consequence of bounded costs or gradient clipping. Clipping bounds a sampled gradient and can introduce bias; the theoretical H is not identified with the experimental threshold 1.0 (Remark A.1). Proposition A.3 gives a separate finite-time bound for clipping the exact-DRO-weighted estimator, including an explicit clipping-error term; it is not a guarantee for the full training implementation.

Proposition 2 (M_ϕ-smoothness). Under Assumptions 2 and 3 (i.e., each L_g is M_L-smooth and $\|\nabla L_g\| \le H$), $\phi(\theta)$ is M_ϕ-smooth with $M_\phi = M_L + GH^2/(2\lambda)$.

The smoothness constant M_ϕ grows with smaller λ (stronger robustness), reflecting the sharpening of the softmax weights. The proof (Appendix A.10) combines the M_L-smoothness of each L_g with the Lipschitz continuity of the softmax map (Lemma A.5).

5.2 Convergence of the idealized algorithm

Assumption 4 (Gradient dispersion). $\frac{1}{G}\sum_{g=1}^{G} \left\|\nabla\phi_g(\theta) - \frac{1}{G}\nabla\phi(\theta)\right\|^2 \le \zeta^2$, where $\nabla\phi_g(\theta) := \xi_g(\theta)\nabla L_g(\theta)$.

Assumption 5 (Bounded local gradient variance). $\mathbb{E}[\|\hat{d}^{(t)}_{g,k} - \nabla\phi_g(\theta^{(t)}_{g,k-1})\|^2 \mid \theta^{(t)}_{g,k-1}] \le \sigma^2$ for every $g \in \mathcal{G}$ and local step k. This bound applies to the exact-DRO-weighted, unclipped estimator.

Assumption 6 (Sampling model). At round t, let $\mathcal{F}^{(t)}$ denote the history before drawing the participating set, so that $\theta^{(t)}$ is $\mathcal{F}^{(t)}$-measurable. Let $\mathcal{H}^{(t)}_{g,k-1}$ denote the history before client g's step-k minibatch; it contains $\mathcal{F}^{(t)}$, $S^{(t)}$, and that client's local iterates and sampling history through step $k-1$.

(i) $|S^{(t)}| = m$ for a fixed $m \in \{1, \dots, G\}$, and $S^{(t)}$ is drawn uniformly without replacement from $\mathcal{G}$, independently of $\mathcal{F}^{(t)}$.

(ii) Each local minibatch is fresh: conditional on $\mathcal{H}^{(t)}_{g,k-1}$, its instances are i.i.d. from $\mathcal{D}_g$ and its trajectories are sampled independently from $\pi_{\theta^{(t)}_{g,k-1}}(\cdot \mid x)$.

(iii) Conditional on $\mathcal{F}^{(t)}$ and $S^{(t)}$, the local sampling randomness of distinct participating clients is independent.

Theorem 1 (Convergence of FedAvg for KL-DRO). Under Assumptions 2–6, let $\Phi_0 = \phi(\theta^{(0)}) - \phi^\star > 0$ and $C_{\mathrm{fed}} = \sigma^2/m + 2H^2$. With $\eta = \sqrt{2\Phi_0/(M_\phi K^2 T C_{\mathrm{fed}})}$ and $T \ge 2M_\phi\Phi_0/C_{\mathrm{fed}}$, so that $K\eta \le 1/M_\phi$, Algorithm 2 achieves

$$\min_{0\le t<T} \mathbb{E}\left[\left\|\nabla\phi(\theta^{(t)})\right\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{10}$$

The expectation includes both client selection and local minibatch randomness. The finite-time bound for any admissible positive η is given in Eq. (41).

Under the stated sampling model, the client-selection deviation has conditional mean zero and contributes to variance. For general K, the remaining client-drift term in Eq. (41) is $\mathcal{O}(\eta^2(K-1)^2)$ and decreases as $\mathcal{O}(1/T)$ under the displayed step size. Setting $K = 1$ eliminates this term; Corollary B.1 gives the corresponding client-sampling variance bound. The rate matches Guo and Yang's FGDRO-KL; see Remark B.4 for a comparison of assumptions.

5.3 Convergence of the EMA-extended algorithm

Theorem 1 assumes exact DRO weights $\xi_g(\theta)$, which require the true expected losses of all clients—unavailable in practice. Algorithm 1 removes this requirement by estimating weights via EMAs, at the cost of additional bias terms in the bound. We now quantify this price; the EMA analysis additionally requires a variance bound on the fresh stochastic gradient before DRO weighting and direction averaging. This is a separate assumption, not a consequence of bounded tour costs.

Assumption 7 (Bounded stochastic gradient variance). $\mathbb{E}[\|\widehat{\nabla L_g} - \nabla L_g(\theta)\|^2 \mid \mathcal{F}] \le \sigma^2$ for all $g \in \mathcal{G}$, where $\mathcal{F}$ is the history before the current minibatch and θ is its $\mathcal{F}$-measurable local iterate.

Theorem 2 (Finite-time descent with EMA tracking). Under Assumptions 1–3 and 7, run Algorithm 1 with full participation ($m = G$), $K = 1$, and the initialization and conditional input moments in Appendix A.13.1. For $\eta > 0$ and $\beta_u, \beta_v, \beta_d \in (0, 1]$ fixed within each run, every integer $T \ge 1$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\phi(\theta^{(t)})\right\|^2\right] \le \frac{2G\Phi_0}{\eta T} + \frac{G^2}{T}\sum_{t=0}^{T-1} Q_t + M_\phi G\eta D_{\mathrm{ema}}^2, \tag{11}$$

where $\Phi_0 = \phi(\theta^{(0)}) - \inf_\theta \phi(\theta)$, $D_{\mathrm{ema}}^2 = e^{4B_L/\lambda}(H^2+\sigma^2)/G^2$, and the deterministic Q_t bounds the expected squared conditional bias (Lemma A.16). The minimum expected squared gradient norm over $0 \le t < T$ obeys the same bound.

The explicit tracking bounds and proof appear in Appendices A.13.3–A.13.6. Theorem 2 alone does not assert that its time-averaged bias term vanishes. Setting all EMA coefficients to 1 does not recover exact weights: the current loss remains noisy and the gradient update still uses the old denominator $v^{(t)}$. No dispersion assumption is used; partial participation and general K require additional retained-state and step-level tracking analysis.

By coupling tracking with descent, Proposition A.5 and Corollary A.1 further give an $\mathcal{O}(T^{-1/2})$ stationarity guarantee for $m = G$, $K = 1$ with $\eta = \eta_0/\sqrt{T}$ and $\beta_u = \beta_v = \beta_d = \beta_0/\sqrt{T}$. Here $\eta_0 > 0$ and $0 < \beta_0 \le 1$ are independent of T and satisfy $c_2(\eta_0/\beta_0)^2 < 1/2$, with c_2 specified in Eq. (56). The parameters are chosen before each run and held fixed during it; this is not a rate claim for arbitrary fixed EMA settings or a within-run decay schedule. Algorithm 1 is unchanged.

6 Experiments

We evaluate our federated DRO framework on synthetic CVRP instances with heterogeneous distribution groups.

6.1 Setup

Distribution groups. We consider G=16 clients, each holding a private CVRP dataset drawn from one of the customer-location distribution types (*cluster*, *explosion*, *grid*) also considered by Jiang et al. [2022], Bossek et al. [2019], with integer demands sampled uniformly from $\{1, \dots, 9\}$ and vehicle capacity 40 for n=50 and 50 for n=100. We vary the distribution pair, problem size ($n \in \{50, 100\}$), minority ratio ($\frac{1}{16}$, $\frac{4}{16}$, or $\frac{8}{16}$), and model dimension ($d \in \{32, 128\}$), yielding twelve conditions (Appendix C). A supplementary setting also includes uniform, implosion, and expansion; all generator procedures and parameters are specified in Appendix C.

Policy network. We use the Attention Model of Kool et al. [2019] with 3 encoder layers. We evaluate two model capacities: a compact model with embedding and hidden dimensions d=32, and a full-size model with d=128. The policy is trained with REINFORCE [Williams, 1992, Sutton and Barto, 2018] using an exponential moving average baseline (β_{bl}=0.8). In the C11 single-rollout implementation, the baseline is updated using the current minibatch and BatchNorm is used in

Table 1: Worst-group cost (↓) across representative conditions. Final-epoch values after 20 epochs, mean ± sample SD across 3 seeds; no statistically significant differences between ERM and our methods (KL/EMA) were found at α=0.05. KL and EMA denote divisive implementations; λ is the nominal parameter, not the time-varying effective temperature. *Minority*: fraction of clients with the underrepresented distribution type; all conditions use *cluster* as the majority type.

		Ours (FedDRO-KL)		FedAvg-GroupDRO	
Condition	FedAvg-ERM	KL λ=2	EMA λ=5	η_q=0.01	η_q=0.001
C1 (expl, n=50, $\frac{1}{16}$)	$20.34_{\pm 1.51}$	$20.30_{\pm 1.61}$	$20.30_{\pm 1.66}$	$24.26_{\pm 0.66}$	$20.29_{\pm 1.65}$
C5 (expl, n=100, $\frac{1}{16}$)	$42.55_{\pm 2.97}$	$42.29_{\pm 2.88}$	$42.50_{\pm 2.94}$	$48.92_{\pm 3.04}$	$42.32_{\pm 3.46}$
C8 (grid, n=50, $\frac{4}{16}$)	$21.31_{\pm 2.00}$	$21.28_{\pm 2.00}$	$21.27_{\pm 2.00}$	$23.97_{\pm 2.17}$	$21.24_{\pm 2.01}$
C11 (grid, n=100, $\frac{1}{16}$)	$48.13_{\pm 0.91}$	$47.59_{\pm 1.01}$	$47.47_{\pm 1.11}$	$55.66_{\pm 0.54}$	$47.71_{\pm 0.83}$

training mode; these implementation details are not covered by the theoretical estimator assumptions (Appendix A.3).

Baselines. We compare four methods: (i) **FedAvg-ERM**, standard federated averaging with uniform weights as in Eq. (2); (ii) **FedAvg-GroupDRO** [Sagawa et al., 2020] with $\eta_q \in \{0.001, 0.005, 0.01\}$; (iii) **FedDRO-KL** (Algorithm 1, β_v=β_d=1); and (iv) **FedDRO-KL-EMA** (Algorithm 1, full three-EMA). See Appendix C for detailed specifications. In experimental results, FedDRO-KL and FedDRO-KL-EMA denote the divisive implementations unless additive centering is explicitly specified.

Hyperparameters. We sweep $\lambda \in \{2, 5, 10\}$ and fix $\beta_u = \beta_v = 0.05$, $\beta_d = 0.1$; all methods use learning rate $\eta = 10^{-3}$ with full client participation each round. Training runs for T=2,000 rounds (20 epochs × 100), 3 seeds; we report mean ± sample standard deviation across seeds for *worst-group cost* (greedy decoding). Optimizer details (momentum, weight decay, clipping) and runtime appear in Appendix A.3.

6.2 Main results

FedDRO-KL matches ERM with no significant degradation. Table 1 summarizes worst-group costs across representative conditions. With 10 random seeds on C11 (cluster/grid, n=100, minority $\frac{1}{16}$, 40 epochs), none of the DRO variants achieves a statistically significant worst-group improvement over ERM (paired t-test, all $p > 0.28$), nor does it cause significant degradation. A six-distribution federated setting (16 clients, λ=5, 10 seeds) confirms this; the upper endpoint of the paired 95% confidence interval is 0.5% of the observed ERM mean (Remark A.3). Crucially, FedDRO-KL preserves average cost within ±0.5% across all conditions. The fixed-temperature KL-DRO objective provides a formal robustness certificate through the worst-group bound (Lemma A.1); its empirical tightness is evaluated in Remark A.2. Results for all twelve conditions, including d=128, appear in Tables 5–6 (Appendix); gaps to HGS reference solutions appear in Table 7.

GroupDRO with aggressive step sizes collapses. FedAvg-GroupDRO with η_q=0.01 degrades worst-group cost across all twelve conditions by 5–26% (Table 6). The exponentiated-gradient update concentrates weight onto the minority client (gradient scale ≈15× for η_q=0.001 at epoch 40 on C11; Figure 4b), effectively discarding all other clients; with η_q=0.01 this collapse is rapid and accompanies worsening worst-group cost (Figure 4a; see also Figure 3 in the Appendix). This supports the theoretical motivation for KL regularization (Section 3.2). A centralized (G=2) baseline confirms that collapse also occurs without federation (Appendix C.5).

Value of federated training. Independently trained minority-client models degrade by 26% in overall validation cost relative to federated models (Appendix C.4); FedDRO-KL preserves this federated advantage with a robustness certificate. This gap is asymmetric: the minority client's independent model specializes on its own distribution (−15% cost) but catastrophically fails on the majority distribution (+31%), whereas the majority client is nearly self-sufficient (< 2% gap), underscoring that federation is most critical for data-scarce participants.

Table 2: Inter-group gradient cosine similarity (minority–majority mean ± sample SD across epochs in each window). Single-rollout: condition C11 (G=16, 40 epochs, seed 42); POMO-100: G=2, epochs 200–249 (post-convergence). Full statistics in Table 4 (Appendix B.4).

Setting	FedAvg-ERM	FedDRO-KL
Single-rollout (λ=2)	-0.007 ± 0.033	$+0.003 \pm 0.030$
POMO-100 (λ=5)	-0.071 ± 0.208	-0.072 ± 0.186

Inter-group gradient orthogonality. To understand why re-weighting consistently fails to improve worst-group cost, we measure inter-group gradient cosine similarity at every epoch (Appendix B.4). The mean minority–majority cosine is near zero for both ERM and FedDRO-KL in the C11 Attention Model runs (Table 2). This finding persists even with $100\times$ variance reduction (POMO-100, Table 2; details in Appendix C.6). These observations are specific to CVRP with the evaluated architecture; whether they extend to scheduling, packing, or other architectures remains open.

EMA approximation matches FedDRO-KL. FedDRO-KL-EMA closely matches FedDRO-KL across all conditions (within one standard deviation; Table 1; full results in Table 6). We recommend reduced FedDRO-KL ($\beta_v = \beta_d = 1$) as the practical default, since the full three-EMA variant offers no measured performance advantage.

Ablation studies (Appendices C.1–C.6) report the effects of temperature λ, local steps K, and C1 partial participation (down to one client); a two-client POMO-100 experiment confirms that the similar DRO–ERM performance, gradient orthogonality, and GroupDRO collapse persist under $100\times$ variance reduction (Appendix C.6).

7 Conclusion

We presented the first federated Group DRO framework for REINFORCE-trained NCO solvers with $\mathcal{O}(1/\sqrt{T})$ convergence under explicit gradient and variance assumptions: uniform partial participation for the idealized algorithm, and full participation with $K = 1$ and horizon-dependent parameters for EMA. KL regularization avoids cumulative weight collapse, providing a formal worst-group robustness certificate, while the evaluated divisive implementation matches ERM quality and preserves the federated advantage for minority clients. Inter-group REINFORCE gradients are near-orthogonal even under $100\times$ variance reduction (Appendix C.6), suggesting that architecture- or representation-level interventions merit investigation in CVRP.

Limitations. Our experiments use synthetic CVRP ($n \leq 100$, G=16); only C1 tests partial participation, down to one client, and real-world and cross-device settings are not evaluated. The guarantees do not directly cover our use of a same-batch EMA baseline, training-mode BatchNorm, clipping, momentum, or weight decay (Appendix A.3). The separate clipped-update bound assumes control of time-averaged clipping error; reported diagnostics are implementation-level proxies. Theorem 1 assumes fixed-size uniform participation and fresh independent local sampling. Its σ^2/m term grows as m decreases and was not estimated in our experiments. EMA analysis covers full participation and $K = 1$; partial participation, general K, and retained-state staleness need further analysis. Corollary A.1 is a horizon-dependent result, not a guarantee for arbitrary fixed coefficients or within-run decay. Smoothness, bounded-gradient, and variance assumptions remain empirically unverified. CVaR, adaptive step sizes [Guo and Yang, 2024], and demand/capacity heterogeneity remain open.

Acknowledgments and Disclosure of Funding

We thank Makoto Nakashima for helpful discussions on the theoretical analysis.

References

D. Angioni, C. Archetti, and M. G. Speranza. Neural combinatorial optimization: A tutorial. *Computers & Operations Research*, 182:107102, 2025.

S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

Y. Bengio, A. Lodi, and A. Prouvost. Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.

J. Bossek, P. Kerschke, A. Neumann, M. Wagner, F. Neumann, and H. Trautmann. Evolving diverse TSP instances by means of novel and creative mutation operators. In *Proceedings of the 15th ACM/SIGEVO Conference on Foundations of Genetic Algorithms*, pp. 58–71, 2019.

J. Bi, Y. Ma, J. Wang, Z. Cao, J. Chen, Y. Sun, and Y. M. Chee. Learning generalizable models for vehicle routing problems via knowledge distillation. In *Advances in Neural Information Processing Systems*, 35, 2022.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Y. Deng, M. M. Kamani, and M. Mahdavi. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems*, 33, 2020.

D. Drakulic, S. Michel, F. Mai, A. Sors, and J.-M. Andreoli. BQ-NCO: Bisimulation quotienting for efficient neural combinatorial optimization. In *Advances in Neural Information Processing Systems*, 36, 2023.

J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

Z. Guo and T. Yang. Communication-efficient federated group distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 37, 2024.

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938, 2018.

A. Hottung, Y.-D. Kwon, and K. Tierney. Efficient active search for combinatorial optimization problems. In *International Conference on Learning Representations*, 2022.

Y. Jiang, Y. Wu, Z. Cao, and J. Zhang. Learning to solve routing problems via distributionally robust optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9786–9794, 2022.

C. K. Joshi, Q. Cappart, L.-M. Rousseau, and T. Laurent. Learning the travelling salesperson problem requires rethinking generalization. *Constraints*, 27:70–98, 2022.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143, 2020.

A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *AISTATS*, pp. 4519–4529, 2020.

S. Khodadadian, P. Sharma, G. Joshi, and S. T. Maguluri. Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*, pp. 10997–11057, 2022.

W. Kool, H. van Hoof, and M. Welling. Attention, learn to solve routing problems! In *International Conference on Learning Representations*, 2019.

D. Kuhn, S. Shafiee, and W. Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

Y.-D. Kwon, J. Choo, B. Kim, I. Yoon, Y. Gwon, and S. Min. POMO: Policy optimization with multiple optima for reinforcement learning. In *Advances in Neural Information Processing Systems*, 33, 2020.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368, 2021.

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2020.

B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 34, 2021.

S. Liu, U. Rückert, and Y. Jin. FOG: A unified framework for federated combinatorial optimization on graphs. In *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2024.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282, 2017.

X. Meng, J. Zhou, J. Gao, Y. Lu, Y. Wu, G. Yuan, and Y. Hou. Enhancing cross-problem vehicle routing via federated learning. *arXiv preprint arXiv:2604.10652*, 2026.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625, 2019.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, 29, 2016.

M. Nazari, A. Oroojlooy, L. V. Snyder, and M. Takáč. Reinforcement learning for solving the vehicle routing problem. In *Advances in Neural Information Processing Systems*, 31, 2018.

H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:4, 2022.

S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, 28, 2015.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 30, 2017.

T. Vidal. Hybrid genetic search for the CVRP: Open-source implementation and SWAP* neighborhood. *Computers & Operations Research*, 140:105643, 2022.

T. Vidal, T. G. Crainic, M. Gendreau, N. Lahrichi, and W. Rei. A hybrid genetic algorithm for multidepot and periodic vehicle routing problems. *Operations Research*, 60(3):611–624, 2012.

O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, 28, 2015.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

B. Woodworth, K. K. Patel, S. U. Stich, Z. Dai, B. Bullins, H. B. McMahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, pp. 10334–10343, 2020.

T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 33, 2020.

L. Zhang, P. Zhao, Z.-H. Zhuang, T. Yang, and Z.-H. Zhou. Stochastic approximation approaches to group distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 36, 2023.

J. Zhou, Z. Cao, Y. Wu, W. Song, Y. Ma, J. Zhang, and C. Xu. MVMoE: Multi-task vehicle routing solver with mixture-of-experts. In *International Conference on Machine Learning*, pp. 61804–61824, 2024.

A Additional Details and Proofs

A.1 Idealized algorithm pseudocode

Algorithm 2 gives the full pseudocode of the idealized FedDRO-KL procedure described in Section 4 (see also Algorithm 1 for the practical EMA variant).

Algorithm 2 FedDRO-KL for NCO (idealized)

Require: Initial $\theta^{(0)}$; learning rate η; local steps K; temperature λ; reference p

for $t = 0, 1, \ldots, T-1$ **do**
  Broadcast $\theta^{(t)}$ to all clients $g \in S^{(t)}$
  for each client $g \in S^{(t)}$ **in parallel do**
    $\theta_{g,0}^{(t)} \leftarrow \theta^{(t)}$
    for $k = 1, \ldots, K$ **do**
      Sample minibatch; compute $\hat{d}_{g,k}^{(t)}$ via Eq. (6)
      $\theta_{g,k}^{(t)} \leftarrow \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}$
    end for
    Send $\theta_{g,K}^{(t)}$ to server
  end for
  $\theta^{(t+1)} \leftarrow \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \theta_{g,K}^{(t)}$
end for

A.2 Notation Summary

We summarize the notation used throughout the paper.

- G: number of clients (groups); index set $\mathcal{G} = \{1, \ldots, G\}$.
- $g \in \mathcal{G}$: client/group index.
- $\mathcal{D}_g$: (unknown) data distribution for client g.
- $x \sim \mathcal{D}_g$: problem instance (e.g., CVRP instance).
- n: number of customers per CVRP instance (e.g., 50 or 100).
- τ: feasible solution (e.g., tour/route); $\ell(\tau) \in \mathbb{R}_+$: solution cost.
- d: embedding and hidden dimension of the policy network (e.g., 32 or 128).
- D: total number of model parameters (dimension of θ; $D \approx 119$K for d=32, $D \approx 693$K for d=128).
- $\pi_\theta(\tau \mid x)$: stochastic policy parameterized by $\theta \in \mathbb{R}^D$.
- $p \in \Delta_G$: reference distribution for KL-regularized DRO.
- $\lambda > 0$: temperature parameter for KL-regularized DRO.
- Client-wise expected loss:

$$L_g(\theta) = \mathbb{E}_{x \sim \mathcal{D}_g}\, \mathbb{E}_{\tau \sim \pi_\theta(\cdot|x)}[\ell(\tau)].$$

- ERM objective:

$$\min_\theta \frac{1}{G} \sum_{g=1}^{G} L_g(\theta).$$

- Group DRO (minimax) objective:

$$\min_\theta \max_{q \in \Delta_G} \sum_{g=1}^{G} q_g L_g(\theta), \qquad \Delta_G = \{q \in \mathbb{R}_+^G : \sum_{g=1}^{G} q_g = 1\}.$$

- KL-regularized Group DRO objective with reference $p \in \Delta_G$ and temperature $\lambda > 0$:

$$\min_{\theta} \left(\max_{q \in \Delta_G} \sum_{g=1}^{G} q_g L_g(\theta) - \lambda \sum_{g=1}^{G} q_g \log \frac{q_g}{p_g} \right).$$

- Smooth (KL-regularized) DRO surrogate:

$$\phi(\theta) = \lambda \log \left(\sum_{g=1}^{G} p_g \exp\left(\frac{L_g(\theta)}{\lambda} \right) \right).$$

- Softmax weights (client importance):

$$\xi_g(\theta) = \frac{p_g \exp(L_g(\theta)/\lambda)}{\sum_{h=1}^{G} p_h \exp(L_h(\theta)/\lambda)}.$$

Federated learning protocol. We refer to the idealized algorithm (exact DRO weights) as **FedDRO-KL** (Algorithm 2) and to the EMA-extended variant as **FedDRO-KL-EMA** (Algorithm 1).

- T: number of communication rounds.
- K: number of local SGD steps per client per round.
- η: local learning rate (step size); effective global step size is $K\eta$.
- $S^{(t)} \subset \mathcal{G}$: set of participating clients at round t. sampled uniformly without replacement under Assumption 6.
- $m = |S^{(t)}|$: fixed number of participating clients.
- B: minibatch size.
- $\bar{\ell}_{g,k}^{(t)}$: batch mean cost for client g at local step k, round t.
- $\widehat{\nabla L_g}$: REINFORCE-based stochastic gradient of L_g from a single minibatch.
- β_{bl}: EMA parameter for the REINFORCE baseline (0.8 in experiments).
- $\theta^{(t)} \in \mathbb{R}^D$: global (broadcast) model at round t.
- $\theta_{g,k}^{(t)}$: local iterate of client g at local step k, round t.
- $\hat{d}_{g,k}^{(t)}$: local stochastic gradient estimator for client g at local step k, round t. In the idealized algorithm (Algorithm 2), this is the per-step REINFORCE gradient scaled by the exact DRO weight; in the EMA algorithm (Algorithm 1), it is the EMA-smoothed gradient direction (see EMA quantities below).
- $\tilde{d}_{g,k}^{(t)}, e_{g,k}^{(t)}$: clipped exact-weight direction and clipping error, respectively, in Proposition A.3.
- $\hat{d}^{(t)} := \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \hat{d}_{g,k}^{(t)}$: aggregated update direction; global update is $\theta^{(t+1)} = \theta^{(t)} - K\eta\, \hat{d}^{(t)}$.
- $\phi^\star := \inf_\theta \phi(\theta)$: optimal DRO objective value.
- $\Phi_0 := \phi(\theta^{(0)}) - \phi^\star$: initial optimality gap.

Gradient-related quantities.

- $\nabla \phi_g(\theta) := \xi_g(\theta) \nabla L_g(\theta)$: g-th weighted component of $\nabla \phi(\theta) = \sum_g \nabla \phi_g(\theta)$.
- $Z(\theta) := \sum_{h=1}^{G} p_h \exp(L_h(\theta)/\lambda)$: partition function.
- $w_g(\theta) := p_g \exp(L_g(\theta)/\lambda)$: unnormalized weight (so $\xi_g = w_g/Z$).
- $v^{(t)}$: server-side normalization constant (approximation to $Z(\theta^{(t)})$).
- $p_{\min} := \min_g p_g$.
- $\mathcal{F}^{(t)}$: history before round-t client selection in Assumption 6.

- $\mathcal{H}_{g,k-1}^{(t)}$: history before client g's fresh step-k minibatch, as defined in Assumption 6.
- $b^{(t)}$: bias of the aggregated direction, with section-specific conditioning. In Lemma A.8, $b^{(t)} = \mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] - G^{-1}\nabla\phi(\theta^{(t)})$; in Lemma A.16, $b^{(t)} = \mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}] - G^{-1}\nabla\phi(\theta^{(t)})$.

Smoothness and bound constants.

- M_L: Lipschitz constant of ∇L_g (Assumption 2).
- H: uniform bound on $\|\nabla L_g(\theta)\|$ (Assumption 3).
- $H_c > 0$: norm-clipping threshold for the exact-weight stochastic direction in Proposition A.3.
- $M_\phi := M_L + GH^2/(2\lambda)$: smoothness constant of ϕ (Proposition 2).
- B_L: uniform bound on cost magnitude, $|\ell(\tau)| \le B_L$ a.s. (Assumption 1).
- ζ^2: gradient dispersion (heterogeneity) across clients (Assumption 4).
- σ^2: bound on stochastic gradient variance (Assumptions 5, 7).
- σ_ℓ^2: bound on instantaneous loss estimation error variance (Lemma A.11 gives the admissible choice $\sigma_\ell^2 = B_L^2/B$).

Convergence rate quantities.

- $C_{\mathrm{fed}} := \sigma^2/m + 2H^2$: second-moment constant in the non-EMA rate (Theorem 1).
- $D_{\mathrm{ema}}^2 = e^{4B_L/\lambda}(H^2+\sigma^2)/G^2$: unconditional second-moment bound for the actual EMA direction with $m = G$ and $K = 1$ (Lemma A.12).
- Q_t: deterministic bound on $\mathbb{E}[\|b^{(t)}\|^2]$ for the history-conditional EMA bias with $m = G$ and $K = 1$ (Lemma A.16).

EMA quantities ($m = G$, $K = 1$; Section A.13).

- $u_{g,1}^{(t)}$: loss EMA estimator (end-of-round state).
- $\hat{w}_{g,1}^{(t)} = p_g \exp(u_{g,1}^{(t)}/\lambda)$: current EMA-based weight.
- $\tilde{v}_{g,1}^{(t)}$: post-update normalization EMA state.
- $\hat{d}_{g,1}^{(t)}$: post-update gradient EMA state; $\hat{d}^{(t)} = G^{-1}\sum_g \hat{d}_{g,1}^{(t)}$ under full participation.
- $\beta_u, \beta_v, \beta_d \in (0,1]$: EMA parameters for loss, normalization, and gradient, respectively.
- $\rho_{\theta,\mathrm{RMS}}$: unconditional RMS model-movement bound (Proposition A.4).
- ϵ_u: unconditional loss EMA tracking bound (Lemma A.13).
- V_t: deterministic bound on $\|v^{(t)} - Z(\theta^{(t)})\|_{L^2}$ (Lemma A.14).
- R_t, B_t: comparison-EMA error and squared conditional-staleness-bias bounds (Lemma A.15).
- A_t: coefficient-error bound used in Lemma A.16.
- $c_g^{(t)} = \hat{w}_{g,1}^{(t)}/v^{(t)} - \xi_g(\theta^{(t)})$: actual-minus-exact weight coefficient.
- $e^{(t)} = \hat{d}^{(t)} - G^{-1}\nabla\phi(\theta^{(t)})$: actual direction error in the coupled analysis.
- $\mathsf{U}_t, \mathsf{N}_t, \mathsf{D}_t, \mathsf{E}_t$: unconditional second moments; their barred versions are time averages, and $\bar{\mathsf{G}}_T$ is the average expected squared gradient norm (Section A.13.7).
- c_0, c_1, c_2: explicit coupled-tracking constants in Proposition A.5.

GroupDRO baseline.

- η_q: step size for the exponentiated-gradient weight update in GroupDRO (Proposition 1).
- $\delta_{\min}$: minimum loss gap between the highest-loss group and all other groups (Proposition 1).

Norm convention. Unless otherwise noted (e.g., $\|\cdot\|_1$ in Section A.9), $\|\cdot\|$ denotes the ℓ_2 (Euclidean) norm.

A.3 Practical Implementation Details

Runtime and communication cost. A full training run (T=2,000 rounds, G=16 clients, d=32, n=100) takes approximately 10.5 hours on a single NVIDIA V100 (32 GB); smaller instances (n=50) finish in roughly half the time. Communication cost per round is one full model broadcast ($\sim$119K parameters for d=32; $\sim$693K for d=128).

The communication estimate above counts model payload only. The analysis protocol in Algorithm 1 broadcasts $v^{(t)}$ and aggregates client estimates $\tilde{v}_g$. In contrast, the reported divisive implementation computes $\bar{u}$ from all retained client loss-EMA states u_g in its single-process training loop, processing clients sequentially; it does not implement a client-to-server exchange of u_g or a separate broadcast of $\bar{u}$. A distributed realization would require state sharing or equivalent server coordination, which is not included in that estimate.

Centered logits. The reported KL experiments use divisive logits $z_g = (u_g/\bar{u} - 1)/\lambda$, where $\bar{u} = (1/G)\sum_h u_h > 0$ for the routing costs considered here. For a fixed set of loss estimates, division by $\bar{u}$ preserves their ordering but changes the ratios of normalized weights relative to logits u_g/λ. It is therefore not merely a numerically stable evaluation of the same fixed-temperature weights. The implementation retains one loss-EMA state u_g per client and computes $\bar{u}$ from all G stored client states whenever a participating client is processed. Under partial participation, nonparticipating clients retain their previous u_g values and thus contribute their most recently stored states. Participating clients are processed sequentially: the current client's u_g is updated before $\bar{u}$ and the logits are recomputed, so later clients in that round use the updates already made to earlier clients' states. The main-results runs use full participation, so nonparticipant staleness does not arise in those rounds.

For the uniform prior, define the normalized weight $q_g = \exp(z_g)/\sum_h \exp(z_h)$. The local gradient scale is $s_g = \exp(z_g)/\operatorname{mean}_h \exp(z_h) = Gq_g$, which has mean one across clients. The code applies this scale to the REINFORCE term before backpropagation, then clips the combined policy gradient norm to 1.0 before updating the gradient EMA. Thus the clipping operation acts on the already scaled gradient.

Effective temperature. Rewriting the centered logit as $z_g = (u_g - \bar{u})/(\lambda\bar{u})$ reveals that this divisive centering is equivalent to the standard additive centering $z'_g = (u_g - \bar{u})/\lambda$ but with an effective temperature $\lambda_{\text{eff}} = \lambda\bar{u}$. For a fixed set of loss estimates and a uniform prior, increasing the temperature moves the normalized weights toward uniformity. Since $\bar{u}$ changes during training, λ_{eff} is data-dependent. The convergence results concern a fixed $\lambda > 0$ and the specified theoretical estimators; this algebraic equivalence does not extend their guarantees to divisive centering or identify the complete server-normalized EMA update with the theoretical update. An alternative implementation using additive centering $z'_g = (u_g - \max_h u_h)/\lambda$ preserves the nominal temperature: the factor $\exp(-\max_h u_h/\lambda)$ cancels between numerator and denominator when normalizing the same current weights, including with a nonuniform prior. This cancellation does not automatically apply when a current numerator and a delayed EMA denominator use different scaling factors. The additive-centering experiment is reported in Section A.3.

Additive-centering ablation. We evaluated additive logits at $\lambda = 2$ on the C11 configuration: cluster/grid, $n = 100$, $d = 32$, $G = 16$, one grid-minority client, and 40 epochs. Table 3 reports seeds 42, 1111, and 2222, together with single-seed divisive and ERM references. Costs are first averaged over the last five epochs of each run. The additive row then reports the mean and sample standard deviation across the three seed-level averages. Reference rows report only the seed-42 means, without uncertainty estimates; they are not a seed-matched three-run comparison.

At the final epoch, the additive runs had maximum normalized weight 0.999975 ± 0.000012 (mean $\pm$ sample SD across seeds), with all three concentrating on the grid-minority client (client 1 in the paper's indexing). The divisive seed-42 reference had last-five-epoch mean maximum weight 0.125900; its final-epoch value was 0.127007. At fixed temperature, normalized exponential weights can be strongly concentrated when group-loss differences are large relative to λ; concentration is not by itself a violation of the fixed-λ formulation or its stationarity results. These results apply to the

Table 3: Additive-centering ablation on the C11 specification (cluster/grid, n=100, d=32, G=16, minority $1/16$, 40 epochs). Costs use last-five-epoch means. The additive row gives mean $\pm$ sample SD over three seeds; reference rows are seed 42 only.

Method	Seeds	Average cost	Worst-group cost
FedDRO-KL (additive, $\lambda = 2$)	3	43.77 ± 1.82	60.30 ± 4.17
FedDRO-KL (divisive, $\lambda = 2$)	1	21.52	37.04
FedAvg-ERM	1	21.50	37.41

stated configuration and temperature; they do not identify the cause of the performance difference or predict outcomes at other temperatures or distributions.

Policy-gradient baseline and scope. For the C11 single-rollout configuration ($B = 128$, $\beta_{\mathrm{bl}} = 0.8$), the training code updates its EMA baseline using the current minibatch mean cost and returns the updated value to form advantages for that same minibatch. The first call initializes the baseline to the current mean; subsequent calls weight the previous value by β_{bl} and the current mean by $1 - \beta_{\mathrm{bl}}$. The baseline therefore depends on the current trajectories and does not satisfy the sufficient condition in Remark A.4. Detaching the baseline from automatic differentiation does not remove this statistical dependence. The implementation uses BatchNorm in training mode: each example's policy can depend on all inputs in the minibatch, rather than only its own x as in Eq. (1). The convergence results apply to the estimators defined in the analysis, not to the full training procedure described here.

Remark A.1 (Gradient clipping and Assumption 3). We clip the policy-gradient norm to 1.0 after client weighting and before the optimizer step, hence before the SGD momentum update. The norm is over all policy parameters, not over individual rollouts. This bounds the gradient passed to the optimizer, not the true gradient $\nabla L_g(\theta)$, and therefore does not enforce Assumption 3 with H=1. Note that bounded costs (Assumption 1) alone do *not* imply bounded gradients: the REINFORCE gradient $\nabla_\theta L_g = \mathbb{E}_{\tau\sim\pi_\theta}[\ell(\tau)\,\nabla_\theta \log \pi_\theta(\tau)]$ can have unbounded norm even when $|\ell(\tau)| \le B_L$, because the score function $\nabla_\theta \log \pi_\theta(\tau)$ is unbounded in general.

Clipping is nonlinear and need not preserve conditional unbiasedness, even if the unclipped estimator satisfies Proposition A.1. Theorems 1 and 2 analyze unclipped estimators. Proposition A.3 gives a separate finite-time bound for clipping the exact-DRO-weighted local estimator, with an explicit squared-excess term and a drift bound for the clipped iterates. It does not extend the EMA guarantee to the clipped implementation. We do not assume that this bias is negligible or confined to early training.

Full-step clipping diagnostics. A diagnostic run with the C11 specification (cluster/grid, $n = 100$, $d = 32$, $G = 16$, minority fraction $1/16$), divisive KL weighting with $\lambda = 2$, seed 42, and $K = 1$ records 64,000 pre-clipping gradient norms over 40 epochs (100 rounds per epoch, 16 clients). All 64,000 local updates exceed the threshold 1.0. The mean squared positive norm excess above this threshold is 1,634,582.6 over the full run and 399,858.4 over the final 20% of updates (12,800 records, the last eight epochs). Since every recorded update is clipped, the averages over activated updates and over all updates coincide.

These are implementation-level diagnostics of the scaled gradient passed to the optimizer. The implementation uses EMA-based divisive weights, with scale Gq_g for a uniform reference, whereas Proposition A.3 clips $\xi_g(\theta)\widehat{\nabla L_g}$ with exact weights. Thus the measured excess-square is a proxy, not an estimate of $\bar{b}^{(t)}_{\mathrm{clip}}$ or a validated value for the theoretical excess term in Eq. (44).

Optimizer. All methods use SGD with momentum 0.9, weight decay 10^{-4}, and learning rate $\eta = 10^{-3}$. Our convergence analysis assumes vanilla SGD local updates without momentum or weight decay; these implementation choices require additional analysis and are not covered by the stated guarantees.

Local steps. For the EMA-extended variant (Algorithm 1), we set K=1 in all main experiments, consistent with the theory where K=1 eliminates client drift. An ablation over $K \in \{1, 2, 3\}$ is reported in Appendix C.2.

A.4 Assumption Summary

For convenience, we list all assumptions used in the convergence analysis. Those marked with ♣ are needed only for the EMA-extended analysis (Section A.13). The estimators in Theorems 1 and 2 are baseline-free and unclipped, with fresh sampling conditional on the history before each minibatch (Fact A.1 and Proposition A.1). Baseline subtraction is allowed under Remark A.4, provided the stated moment bounds also hold for that estimator.

#	Name	Used in
2	Smoothness of client losses	Prop. 2, Thm 1, 2
3	Bounded gradients	Prop. 2, A.4, Thm 1, 2
4	Gradient dispersion (heterogeneity ζ^2)	Thm 1
5	Bounded local gradient variance	Thm 1
6	Fixed-size uniform client sampling	Thm 1, Cor. B.1
1♣	Bounded cost	Prop. A.4, Thm 2
7♣	Bounded stochastic gradient variance	Thm 2

Derived properties. The following are not assumptions but consequences derived in the appendix:

- **Proposition A.1** (conditionally unbiased local gradients) — used in Thm 1.
- **Proposition A.2** (history-conditional local second moments) — used in Lemmas A.9 and A.10.
- **Proposition A.4♣** (unconditional RMS model and loss drift bounds) — used in Thm 2.
- **Lemma A.11♣** (conditional cost estimation error, at most B_L^2/B) — used in Thm 2.
- **Lemma A.12♣** (unconditional second moment of the actual EMA direction) — used in Proposition A.4, Theorem 2, and Corollary A.1.

A.5 Smooth DRO surrogate

The KL-regularized simplex maximization has the following softmax solution and log-sum-exp optimal value. For the uniform reference distribution, the equivalent log-sum-exp formulation is used in [Guo and Yang, 2024]. We derive the general-p form below.

KL-regularized Group DRO and its Smooth Surrogate. Let there be G groups (clients), indexed by $g \in \{1, \dots, G\}$, and let $L_g(\theta)$ denote the (group-wise) expected loss under parameter $\theta \in \mathbb{R}^D$. Consider the KL-regularized Group DRO objective with reference $p \in \Delta_G$:

$$\min_{\theta \in \mathbb{R}^D} \left(\max_{q \in \Delta_G} \sum_{g=1}^{G} q_g \, L_g(\theta) \; - \; \lambda \sum_{g=1}^{G} q_g \log \frac{q_g}{p_g} \right), \tag{12}$$

where $\lambda > 0$ and $\Delta_G := \{q \in \mathbb{R}^G : \; q_g \geq 0, \; \sum_{g=1}^{G} q_g = 1\}$.

Fix θ and define the inner maximization problem

$$\max_{q \in \Delta_G} \; F_\theta(q) \quad \text{with} \quad F_\theta(q) := \sum_{g=1}^{G} q_g \, L_g(\theta) - \lambda \sum_{g=1}^{G} q_g \log \frac{q_g}{p_g}. \tag{13}$$

Since the negative entropy term is strictly concave in q, F_θ is strictly concave, so the maximizer is unique.

Lagrangian and stationarity. Introduce a Lagrange multiplier $\nu \in \mathbb{R}$ for the simplex constraint $\sum_{g=1}^{G} q_g = 1$. The Lagrangian is

$$\mathcal{L}(q, \nu) = \sum_{g=1}^{G} q_g \, L_g(\theta) - \lambda \sum_{g=1}^{G} q_g \log \frac{q_g}{p_g} + \nu \left(\sum_{g=1}^{G} q_g - 1 \right). \tag{14}$$

For any optimal solution $(q^\star, \nu^\star)$, the first-order optimality condition yields, for each g,

$$0 = \frac{\partial \mathcal{L}}{\partial q_g}(q^\star, \nu^\star) = L_g(\theta) - \lambda\Big(\log \frac{q_g^\star}{p_g} + 1\Big) + \nu^\star. \tag{15}$$

Rearranging Eq. (15) gives

$$\log \frac{q_g^\star}{p_g} = \frac{L_g(\theta) + \nu^\star - \lambda}{\lambda}, \qquad \text{hence} \qquad q_g^\star = p_g \exp\!\left(\frac{L_g(\theta)}{\lambda}\right) \exp\!\left(\frac{\nu^\star - \lambda}{\lambda}\right). \tag{16}$$

Normalization and the softmax form. Imposing $\sum_{g=1}^G q_g^\star = 1$ in Eq. (16) yields

$$1 = \sum_{g=1}^G q_g^\star = \sum_{g=1}^G p_g \exp\left(\frac{L_g(\theta)}{\lambda}\right) \exp\left(\frac{\nu^\star - \lambda}{\lambda}\right), \tag{17}$$

so

$$\exp\left(\frac{\nu^\star - \lambda}{\lambda}\right) = \frac{1}{\sum_{h=1}^G p_h \exp\left(\frac{L_h(\theta)}{\lambda}\right)}. \tag{18}$$

Substituting Eq. (18) back into Eq. (16) gives the closed form

$$q_g^\star(\theta) = \frac{p_g \exp\left(L_g(\theta)/\lambda\right)}{\sum_{h=1}^G p_h \exp\left(L_h(\theta)/\lambda\right)}, \qquad g = 1, \ldots, G, \tag{19}$$

i.e., the optimal adversarial weights are a softmax over group losses.

Optimal value of the inner problem. Let $Z(\theta) := \sum_{h=1}^G p_h \exp(L_h(\theta)/\lambda)$. From Eq. (19),

$$\log \frac{q_g^\star}{p_g} = \log\left(\frac{p_g \exp(L_g(\theta)/\lambda)}{Z(\theta)} \cdot \frac{1}{p_g}\right) = \frac{L_g(\theta)}{\lambda} - \log Z(\theta).$$

Therefore,

$$\sum_{g=1}^G q_g^\star \log \frac{q_g^\star}{p_g} = \sum_{g=1}^G q_g^\star \left(\frac{L_g(\theta)}{\lambda} - \log Z(\theta)\right) \tag{20}$$

$$= \frac{1}{\lambda} \sum_{g=1}^G q_g^\star L_g(\theta) - \log Z(\theta), \tag{21}$$

since $\sum_g q_g^\star = 1$. Plugging Eq. (21) into Eq. (13) yields

$$F_\theta(q^\star) = \sum_{g=1}^G q_g^\star L_g(\theta) - \lambda \left[\frac{1}{\lambda} \sum_{g=1}^G q_g^\star L_g(\theta) - \log Z(\theta)\right] \tag{22}$$

$$= \lambda \log Z(\theta) = \lambda \log\left(\sum_{g=1}^G p_g \exp\left(\frac{L_g(\theta)}{\lambda}\right)\right). \tag{23}$$

Smooth (KL-regularized) DRO surrogate. Combining Eqs. (12) and (23), the KL-regularized min–max problem is equivalent to minimizing the following smooth surrogate objective:

$$\phi(\theta) := \lambda \log\left(\sum_{g=1}^G p_g \exp\left(\frac{L_g(\theta)}{\lambda}\right)\right). \tag{24}$$

Moreover, the maximizer in the inner problem is given by the softmax weights in Eq. (19). We denote $\xi_g(\theta) := q_g^\star(\theta)$ for brevity.

A.6 Worst-Group Loss Upper Bound via $\phi(\theta)$

We provide a simple upper bound relating the worst-group loss to the smooth DRO objective.

Lemma A.1 (Worst-group bound). For $\phi(\theta) = \lambda \log\Big(\sum_{g=1}^{G} p_g \exp(L_g(\theta)/\lambda)\Big)$,

$$\max_g L_g(\theta) \le \phi(\theta) - \lambda \log\left(\min_g p_g\right).$$

In particular, for uniform $p_g = 1/G$,

$$\max_g L_g(\theta) \le \phi(\theta) + \lambda \log G.$$

Proof. Let $g^\star \in \arg\max_g L_g(\theta)$. Then

$$\sum_{g=1}^{G} p_g \exp(L_g(\theta)/\lambda) \ge p_{g^\star} \exp(L_{g^\star}(\theta)/\lambda) \ge (\min_g p_g) \exp(\max_g L_g(\theta)/\lambda).$$

Taking $\lambda \log(\cdot)$ of both sides yields $\phi(\theta) \ge \lambda \log(\min_g p_g) + \max_g L_g(\theta)$, and rearranging gives the claim. □

Remark A.2 (Numerical tightness of the worst-group bound). With uniform prior $p_g = 1/G$, the bound reads $\max_g L_g(\theta) \le \phi(\theta) + \lambda \log G$. The additive slack $\lambda \log G$ grows with λ, but the surrogate $\phi(\theta)$ simultaneously approaches the average $\bar{L}(\theta) = \sum_g p_g L_g(\theta)$ for large λ. The table evaluates this bound on the final-epoch greedy-validation costs for C1 and C11 using the divisive KL runs and nominal λ. This empirical calculation is neither the divisive training objective nor a bound on population expected losses under stochastic decoding.

Condition	λ	Worst-client cost	Empirical bound	Gap (relative)
C1	2	20.30 ± 1.61	20.55 ± 1.53	0.25 (1.3%)
C1	5	20.27 ± 1.62	26.15 ± 1.08	5.88 (29.0%)
C1	10	20.32 ± 1.57	39.46 ± 0.98	19.14 (94.2%)
C11	2	47.59 ± 1.01	47.59 ± 1.01	0.00 (0.0%)
C11	5	47.90 ± 0.60	48.69 ± 0.81	0.79 (1.6%)
C11	10	47.85 ± 0.78	57.28 ± 1.79	9.43 (19.7%)

At λ=2, the relative gap is approximately 1.3% for C1 and rounds to 0.0% for C11; the latter is not exact equality. As λ increases, the bound becomes progressively looser, consistent with the smoothing interpretation of ϕ as a soft maximum.

Remark A.3 (Effect size and confidence intervals for DRO vs. ERM). To guard against interpreting non-significance as equivalence, we report 95% confidence intervals and Cohen's d for the paired worst-group cost differences (DRO − ERM) across the three main 10-seed comparisons. All use greedy-validation costs at the final epoch (epoch 39), pairing the same ten seeds between methods. Each 95% Student-t interval uses nine degrees of freedom and the sample SD of the seed-level differences; Cohen's d is the mean difference divided by that sample SD.

Condition	Variant	Δ_{worst}	95% CI	Cohen's d
C11 (cluster/grid)	KL2	$+0.04$	$[-0.20, +0.27]$	$+0.11$
C11 (cluster/grid)	KL5	$+0.14$	$[-0.22, +0.51]$	$+0.28$
6-dist (n=100, G=16)	KL5	-0.12	$[-0.43, +0.18]$	-0.29

All effect sizes fall in the "small" range ($|d| < 0.3$) and all confidence intervals contain zero. Their upper endpoints, divided by the corresponding observed ERM means using unrounded values, are 0.64%, 1.23%, and 0.49%, respectively. These results are consistent with a true effect near zero.

A.7 Policy Gradient Estimator (REINFORCE)

We use the standard REINFORCE estimator [Williams, 1992] to obtain an unbiased estimate of $\nabla_\theta L_g(\theta)$.

Fact A.1 (Unbiased REINFORCE estimator; Williams 1992). Let $\mathcal{F}$ denote the history before sampling, with θ measurable with respect to $\mathcal{F}$. For a fixed client g, conditionally on $\mathcal{F}$, sample $x \sim \mathcal{D}_g$ and then $\tau \sim \pi_\theta(\cdot \mid x)$. Assume that the policy has parameter-independent support and that differentiation can be interchanged with expectation. Then

$$\hat{d}_g(\theta; x, \tau) := \ell(\tau) \nabla_\theta \log \pi_\theta(\tau \mid x)$$

satisfies $\mathbb{E}[\hat{d}_g(\theta; x, \tau) \mid \mathcal{F}] = \nabla_\theta L_g(\theta)$. This follows from the log-derivative trick [see, e.g., Sutton and Barto, 2018, Section 13.3].

Remark A.4 (Baseline subtraction). A baseline may be used for variance reduction, but an arbitrary baseline need not reduce variance or preserve unbiasedness. A sufficient condition for unbiasedness is that the baseline $b(x, \mathcal{F})$ is fixed given x and the pre-sampling history $\mathcal{F}$, and does not depend on the current sampled trajectory. Under the regularity conditions of Fact A.1 and integrability of the baseline-score product,

$$\mathbb{E}[b(x, \mathcal{F}) \nabla_\theta \log \pi_\theta(\tau \mid x) \mid x, \mathcal{F}] = b(x, \mathcal{F})\, \mathbb{E}[\nabla_\theta \log \pi_\theta(\tau \mid x) \mid x, \mathcal{F}] = 0.$$

Thus replacing $\ell(\tau)$ by $\ell(\tau) - b(x, \mathcal{F})$ preserves conditional unbiasedness.

Remark A.5 (REINFORCE assumption regime vs. supervised SGD). Adapting the compositional FGDRO framework of Guo and Yang [2024] from supervised SGD to REINFORCE policy gradients is not a straightforward substitution; it induces a qualitatively different *assumption regime*. We summarize the key differences below.

1. **Distinct roles of cost and gradient assumptions.** Guo and Yang's Assumption 5.1 requires each local loss to be Lipschitz continuous and smooth. Our analysis instead states smoothness of the expected losses, bounded true gradients, and stochastic-gradient variance bounds separately. Bounded tour costs (Assumption 1: $|\ell(\tau)| \leq B_L$), natural for routing on bounded instance domains with fixed problem size, are additionally used to control loss and weight tracking.
2. **Bounded costs alone do not imply finite gradient variance.** For the single-sample estimator in Fact A.1,
$$\mathbb{E}\Big[\|\hat{d}_g(\theta; x, \tau)\|^2 \mid \mathcal{F}\Big] \leq B_L^2\, \mathbb{E}\big[\|\nabla_\theta \log \pi_\theta(\tau \mid x)\|^2 \;\big|\; \mathcal{F}\big]\,.$$
Bounded costs control only the first factor on the right-hand side; they do not bound the score-function second moment. We therefore impose Assumption 5 on the exact-weight local estimator and Assumption 7 on the unweighted estimator used as input to the EMA update. The exact weight is fixed conditional on the local sampling history, whereas an EMA weight computed from the current costs can be correlated with the current gradient estimate.
3. **Bounded cost does not imply bounded gradient.** The expected gradient satisfies $\|\nabla L_g(\theta)\| \leq B_L \cdot \mathbb{E}[\|\nabla_\theta \log \pi_\theta(\tau \mid x)\|]$, but the score function norm $\|\nabla_\theta \log \pi_\theta\|$ depends on the full network architecture (the chain rule through a multi-layer transformer produces a product of layer Jacobians whose norm can grow with depth and parameter magnitudes) and is not bounded without additional architectural assumptions. Assumption 3 ($\|\nabla L_g(\theta)\| \leq H$) is therefore an *independent* condition that cannot be derived from Assumption 1 alone.
4. **Bounded costs as essential for EMA tracking.** The bounded-cost assumption also plays a distinct role in Algorithm 1: the EMA state u_g tracks the running average of batch costs $\bar{\ell}_g$, and because $|\bar{\ell}_g| \leq B_L$ (Assumption 1), initialization in $[-B_L, B_L]$ ensures that u_g remains in this interval by convexity, which bounds the exponential weights by $e^{-B_L/\lambda} \leq \exp(u_g/\lambda) \leq e^{B_L/\lambda}$. This propagates through the drift bounds (Proposition A.4) and is essential for controlling the EMA tracking error in Theorem 2. Bounded costs also yield the loss-estimation variance bound under the minibatch sampling conditions of Lemma A.11.

A.8 Gradient of the Smooth DRO Objective

We derive the gradient of $\phi(\theta)$ and show it is a convex combination of client gradients with softmax weights.

Lemma A.2 (Gradient of ϕ). Let $\phi(\theta) = \lambda \log\Big(\sum_{g=1}^{G} p_g \exp(L_g(\theta)/\lambda)\Big)$. Then

$$\nabla_\theta \phi(\theta) = \sum_{g=1}^{G} \xi_g(\theta)\, \nabla_\theta L_g(\theta),$$

where $\xi_g(\theta)$ is defined in Section A.5.

Proof. Let $Z(\theta) = \sum_{g=1}^{G} p_g \exp(L_g(\theta)/\lambda)$. Then

$$\nabla_\theta \phi(\theta) = \lambda \nabla_\theta \log Z(\theta) = \lambda \frac{1}{Z(\theta)} \nabla_\theta Z(\theta).$$

Moreover,

$$\nabla_\theta Z(\theta) = \sum_{g=1}^{G} p_g \exp(L_g(\theta)/\lambda) \cdot \frac{1}{\lambda} \nabla_\theta L_g(\theta).$$

Combining yields

$$\nabla_\theta \phi(\theta) = \sum_{g=1}^{G} \frac{p_g \exp(L_g(\theta)/\lambda)}{\sum_{h=1}^{G} p_h \exp(L_h(\theta)/\lambda)} \nabla_\theta L_g(\theta) = \sum_{g=1}^{G} \xi_g(\theta) \nabla_\theta L_g(\theta).$$

□

A.9 Properties of the Softmax Map

We state some useful properties of the softmax map, which are used in the smoothness analysis of ϕ.

Lemma A.3 (Integral (mean-value) representation for vector-valued maps). Let $U \subset \mathbb{R}^G$ be an open set and let $\tilde{\xi} : U \to \mathbb{R}^G$ be continuously differentiable. Fix $y, z \in U$ and assume the line segment

$$\{z + t(y - z) : t \in [0, 1]\} \subset U.$$

Then

$$\tilde{\xi}(y) - \tilde{\xi}(z) = \int_0^1 J_{\tilde{\xi}}\big(z + t(y - z)\big)(y - z) \mathrm{d}t,$$

where $J_{\tilde{\xi}}(\cdot)$ denotes the Jacobian matrix of $\tilde{\xi}$.

Proof. Define $\gamma : [0, 1] \to U$ by $\gamma(t) := z + t(y - z)$. By assumption, $\gamma([0, 1]) \subset U$. Let $F : [0, 1] \to \mathbb{R}^G$ be the composition $F(t) := \tilde{\xi}(\gamma(t))$. Since γ is C^∞ and $\tilde{\xi}$ is C^1, the composition F is C^1. Applying the fundamental theorem of calculus componentwise to the vector-valued function F gives

$$\tilde{\xi}(y) - \tilde{\xi}(z) = F(1) - F(0) = \int_0^1 F'(t) \mathrm{d}t.$$

By the chain rule,

$$F'(t) = J_{\tilde{\xi}}(\gamma(t))\gamma'(t) = J_{\tilde{\xi}}(z + t(y - z))(y - z),$$

because $\gamma'(t) = y - z$. Substituting this expression for $F'(t)$ into the integral yields the claim. □

Lemma A.4 (Jacobian of the softmax map and an induced-norm bound). Let $p = (p_1, \ldots, p_G) \in \Delta_G$, $\lambda > 0$ and $y = (y_1, \ldots, y_G) \in \mathbb{R}^G$. Define the (temperature-λ) softmax map $\tilde{\xi} : \mathbb{R}^G \to \Delta_G$, $\tilde{\xi} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_G)$ by

$$\tilde{\xi}_g(y) := \frac{p_g \exp(y_g/\lambda)}{\sum_{h=1}^{G} p_h \exp(y_h/\lambda)} \qquad (g = 1, \ldots, G).$$

Then $\tilde{\xi}$ is continuously differentiable and its Jacobian satisfies

$$J_{\tilde{\xi}}(y) = \frac{1}{\lambda}\Big(\mathrm{Diag}(\tilde{\xi}(y)) - \tilde{\xi}(y)\tilde{\xi}(y)^\top\Big).$$

Moreover, the induced ℓ_1-operator norm is bounded as

$$\|J_{\tilde{\xi}}(y)\|_1 \le \frac{1}{2\lambda} \qquad \text{for all } y \in \mathbb{R}^G.$$

(In particular, the same bound holds for $\|J_{\tilde{\xi}}(y)\|_\infty$ by symmetry.)

Proof. Let

$$\tilde{Z}(y) := \sum_{h=1}^{G} p_h \exp(y_h/\lambda), \qquad \text{so that} \qquad \tilde{\xi}_g(y) = \frac{p_g \exp(y_g/\lambda)}{\tilde{Z}(y)}.$$

Differentiate the quotient. For $g, h \in \{1, \dots, G\}$,

$$\frac{\partial \tilde{\xi}_g}{\partial y_h}(y) = \frac{\partial}{\partial y_h}\big(p_g e^{y_g/\lambda}\big)\frac{1}{\tilde{Z}(y)} - \frac{p_g e^{y_g/\lambda}}{\tilde{Z}(y)^2}\frac{\partial \tilde{Z}}{\partial y_h}(y).$$

Since $\frac{\partial}{\partial y_h}\big(p_g e^{y_g/\lambda}\big) = \frac{1}{\lambda} p_g e^{y_g/\lambda}\mathbf{1}\{g = h\}$ and $\frac{\partial \tilde{Z}}{\partial y_h}(y) = \frac{1}{\lambda} p_h e^{y_h/\lambda}$, we obtain

$$\frac{\partial \tilde{\xi}_g}{\partial y_h}(y) = \frac{1}{\lambda}\left(\tilde{\xi}_g(y)\mathbf{1}\{g = h\} - \tilde{\xi}_g(y)\tilde{\xi}_h(y)\right).$$

Stacking these partial derivatives gives the matrix identity

$$J_{\tilde{\xi}}(y) = \frac{1}{\lambda}\big(\mathrm{Diag}(\tilde{\xi}(y)) - \tilde{\xi}(y)\tilde{\xi}(y)^\top\big).$$

For the induced ℓ_1-operator norm, recall $\|A\|_1 = \max_{1 \le j \le G} \sum_{i=1}^{G} |A_{ij}|$. Using the above formula, the (g, h) entry of $J_{\tilde{\xi}}(y)$ equals

$$(J_{\tilde{\xi}}(y))_{gh} = \frac{1}{\lambda}\tilde{\xi}_g(y)\big(\mathbf{1}\{g = h\} - \tilde{\xi}_h(y)\big).$$

Fix a column h. Then

$$\begin{aligned}
\sum_{g=1}^{G} |(J_{\tilde{\xi}}(y))_{gh}| &= \frac{1}{\lambda}\left(|\tilde{\xi}_h(1 - \tilde{\xi}_h)| + \sum_{g \neq h} \tilde{\xi}_g \tilde{\xi}_h\right) = \frac{1}{\lambda}\left(\tilde{\xi}_h(1 - \tilde{\xi}_h) + \tilde{\xi}_h \sum_{g \neq h} \tilde{\xi}_g\right) \\
&= \frac{1}{\lambda}\left(\tilde{\xi}_h(1 - \tilde{\xi}_h) + \tilde{\xi}_h(1 - \tilde{\xi}_h)\right) = \frac{2}{\lambda}\tilde{\xi}_h(1 - \tilde{\xi}_h) \\
&= \frac{1}{\lambda}\left(-2\Big(\tilde{\xi}_h - \frac{1}{2}\Big)^2 + \frac{1}{2}\right) \le \frac{1}{2\lambda},
\end{aligned}$$

where we used $\sum_g \tilde{\xi}_g = 1$. Taking the maximum over h yields $\|J_{\tilde{\xi}}(y)\|_1 \le 1/(2\lambda)$. □

Lemma A.5 (Lipschitz continuity of the softmax map). Let $\tilde{\xi}$ be defined as in Lemma A.4. Then $\tilde{\xi}$ is Lipschitz continuous with Lipschitz constant $1/(2\lambda)$, i.e.,

$$\|\tilde{\xi}(y_1) - \tilde{\xi}(y_2)\|_1 \le \frac{1}{2\lambda}\|y_1 - y_2\|_1 \qquad \text{for all } y_1, y_2 \in \mathbb{R}^G.$$

Proof. By Lemma A.3 and Lemma A.4,

$$\begin{aligned}
\|\tilde{\xi}(y_1) - \tilde{\xi}(y_2)\|_1 &= \left\| \int_0^1 J_{\tilde{\xi}}(y_2 + t(y_1 - y_2))(y_1 - y_2)\,\mathrm{d}t \right\|_1 \\
&\le \int_0^1 \|J_{\tilde{\xi}}(y_2 + t(y_1 - y_2))(y_1 - y_2)\|_1\,\mathrm{d}t \\
&\le \int_0^1 \|J_{\tilde{\xi}}(y_2 + t(y_1 - y_2))\|_1 \cdot \|y_1 - y_2\|_1\,\mathrm{d}t \\
&\le \left(\sup_{t \in [0,1]} \|J_{\tilde{\xi}}(y_2 + t(y_1 - y_2))\|_1\right)\|y_1 - y_2\|_1 \int_0^1 \mathrm{d}t \\
&\le \frac{1}{2\lambda}\|y_1 - y_2\|_1.
\end{aligned}$$

□

A.10 Smoothness of the Smooth DRO Objective

We prove that the KL-regularized DRO surrogate ϕ inherits smoothness from the client losses (Assumptions 2 and 3).

Proof of Proposition 2. We write

$$\begin{aligned}\nabla_\theta\phi(\theta_1) - \nabla_\theta\phi(\theta_2) &= \sum_g \xi_g(\theta_1)\nabla_\theta L_g(\theta_1) - \sum_g \xi_g(\theta_2)\nabla_\theta L_g(\theta_2) \\ &= \underbrace{\sum_g (\xi_g(\theta_1) - \xi_g(\theta_2))\nabla_\theta L_g(\theta_1)}_{\text{(I)}} + \underbrace{\sum_g \xi_g(\theta_2)(\nabla_\theta L_g(\theta_1) - \nabla_\theta L_g(\theta_2))}_{\text{(II)}}.\end{aligned} \tag{25}$$

Term (II). By the M_L-smoothness of L_g (Assumptions 2) and $\sum_g \xi_g(\theta_2) = 1$,

$$\|\text{(II)}\| \le \sum_g \xi_g(\theta_2)\|\nabla_\theta L_g(\theta_1) - \nabla_\theta L_g(\theta_2)\| \le M_L\|\theta_1 - \theta_2\|.$$

Term (I). By the triangle inequality and Assumption 3,

$$\|\text{(I)}\| \le \sum_g |\xi_g(\theta_1) - \xi_g(\theta_2)|\,\|\nabla_\theta L_g(\theta_1)\| \le H\,\|\xi(\theta_1) - \xi(\theta_2)\|_1.$$

From Lemma A.5 (applied in the ℓ_1 norm), substituting $\tilde{\xi}(y) = \tilde{\xi}(L(\theta_1)) := \xi(\theta_1)$, $\tilde{\xi}(z) = \tilde{\xi}(L(\theta_2)) := \xi(\theta_2)$, and $L(\theta) = (L_1(\theta), \dots, L_G(\theta)) \in \mathbb{R}^G$:

$$\|\xi(\theta_1) - \xi(\theta_2)\|_1 \le \frac{1}{2\lambda}\|L(\theta_1) - L(\theta_2)\|_1.$$

By the mean value theorem, there exists $c \in (0,1)$ and $\tilde{\theta} = c\theta_1 + (1-c)\theta_2$ such that

$$L_g(\theta_1) - L_g(\theta_2) = \nabla_\theta L_g(\tilde{\theta})^\top(\theta_1 - \theta_2), \quad g \in \mathcal{G}.$$

Appling the Cauchy–Schwarz inequality and Assumption 3,

$$\begin{aligned}|L_g(\theta_1) - L_g(\theta_2)| &= |\nabla_\theta L_g(\tilde{\theta})^\top(\theta_1 - \theta_2)| \\ &\le \|\nabla_\theta L_g(\tilde{\theta})\|\|\theta_1 - \theta_2\| \\ &\le H\|\theta_1 - \theta_2\|, \quad g \in \mathcal{G},\end{aligned}$$

so $\|L(\theta_1) - L(\theta_2)\|_1 \le GH\|\theta_1 - \theta_2\|$. Therefore

$$\|\text{(I)}\| \le \frac{GH^2}{2\lambda}\|\theta_1 - \theta_2\|.$$

Combining the two bounds and the triangle inequality yields

$$\|\nabla_\theta\phi(\theta_1) - \nabla_\theta\phi(\theta_2)\| \le \left(M_L + \frac{GH^2}{2\lambda}\right)\|\theta_1 - \theta_2\|.$$

□

Remark A.6 (Smoothness without bounded costs). The smoothness of ϕ requires only bounded gradients (Assumption 3) and smoothness of the client losses (Assumption 2); the bounded-cost assumption (Assumption 1) is not needed here, nor in Theorem 1. It enters only in the EMA tracking bounds of Theorem 2.

A.11 One-Step Expected Decrease

We provide a standard descent lemma for stochastic gradients applied to ϕ.

Lemma A.6 (Descent lemma for ϕ). Suppose ϕ is M_ϕ-smooth. For all $\theta_1, \theta_2 \in \mathbb{R}^D$,

$$\phi(\theta_2) \le \phi(\theta_1) + \langle \nabla\phi(\theta_1), \theta_2 - \theta_1 \rangle + \frac{M_\phi}{2}\|\theta_2 - \theta_1\|^2.$$

Proof.

$$\begin{aligned}
\phi(\theta_2) - \phi(\theta_1) &= [\phi(\theta_1 + t(\theta_2 - \theta_1))]_0^1 \\
&= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\phi(\theta_1 + t(\theta_2 - \theta_1))\mathrm{d}t \\
&= \int_0^1 \nabla\phi(\theta_1 + t(\theta_2 - \theta_1))^\top(\theta_2 - \theta_1)\mathrm{d}t \\
&= \int_0^1 \langle \nabla\phi(\theta_1 + t(\theta_2 - \theta_1)), \theta_2 - \theta_1 \rangle \mathrm{d}t \\
&= \int_0^1 \left\langle \left(\nabla\phi(\theta_1) + \left(\nabla\phi(\theta_1 + t(\theta_2 - \theta_1)) - \nabla\phi(\theta_1)\right)\right), \theta_2 - \theta_1 \right\rangle \mathrm{d}t \\
&= \langle \nabla\phi(\theta_1), \theta_2 - \theta_1 \rangle + \int_0^1 \underbrace{\langle \nabla\phi(\theta_1 + t(\theta_2 - \theta_1)) - \nabla\phi(\theta_1), \theta_2 - \theta_1 \rangle}_{\text{(I)}} \mathrm{d}t.
\end{aligned}$$

By the Cauchy–Schwarz inequality and the M_ϕ-smoothness (Proposition 2),

$$\begin{aligned}
|\text{(I)}| &\le \|\nabla\phi(\theta_1 + t(\theta_2 - \theta_1)) - \nabla\phi(\theta_1)\| \cdot t\|\theta_2 - \theta_1\| \\
&\le M_\phi t\|\theta_2 - \theta_1\|^2.
\end{aligned}$$

Finally we obtain

$$\begin{aligned}
\phi(\theta_2) - \phi(\theta_1) &\le \langle \nabla\phi(\theta_1), \theta_2 - \theta_1 \rangle + M_\phi\|\theta_2 - \theta_1\|^2 \int_0^1 t\mathrm{d}t \\
&= \langle \nabla\phi(\theta_1), \theta_2 - \theta_1 \rangle + \frac{M_\phi}{2}\|\theta_2 - \theta_1\|^2.
\end{aligned}$$

□

Lemma A.7 (One-step progress bound). Suppose ϕ is M_ϕ-smooth, $\theta_{t+1} = \theta_t - \eta_t\hat{d}_t$ and $\eta_t > 0$, where $\hat{d}_t \in \mathbb{R}^D$ satisfies

$$\mathbb{E}[\hat{d}_t \mid \theta_t] = \nabla\phi(\theta_t) + b_t \quad \text{and} \quad \mathbb{E}[\|\hat{d}_t\|^2 \mid \theta_t] \le C$$

for some bias $b_t \in \mathbb{R}^D$ and constant $C > 0$. Then

$$\mathbb{E}[\phi(\theta_{t+1}) \mid \theta_t] \le \phi(\theta_t) - \eta_t\|\nabla\phi(\theta_t)\|^2 - \eta_t\langle \nabla\phi(\theta_t), b_t \rangle + \frac{M_\phi}{2}\eta_t^2 C.$$

Proof. By Lemma A.6 (descent lemma),

$$\phi(\theta_{t+1}) \le \phi(\theta_t) + \langle \nabla\phi(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{M_\phi}{2}\|\theta_{t+1} - \theta_t\|^2.$$

Substitute $\theta_{t+1} - \theta_t = -\eta_t\hat{d}_t$ and take conditional expectation,

$$\begin{aligned}
\mathbb{E}[\phi(\theta_{t+1}) \mid \theta_t] &\le \phi(\theta_t) + \mathbb{E}[\langle \nabla\phi(\theta_t), -\eta_t\hat{d}_t \rangle \mid \theta_t] + \frac{M_\phi}{2}\mathbb{E}[\| - \eta_t\hat{d}_t\|^2 \mid \theta_t] \\
&= \phi(\theta_t) - \eta_t\langle \nabla\phi(\theta_t), \mathbb{E}[\hat{d}_t \mid \theta_t] \rangle + \frac{M_\phi}{2}\eta_t^2\mathbb{E}[\|\hat{d}_t\|^2 \mid \theta_t] \\
&\le \phi(\theta_t) - \eta_t\langle \nabla\phi(\theta_t), \nabla\phi(\theta_t) + b_t \rangle + \frac{M_\phi}{2}\eta_t^2 C \\
&= \phi(\theta_t) - \eta_t\|\nabla\phi(\theta_t)\|^2 - \eta_t\langle \nabla\phi(\theta_t), b_t \rangle + \frac{M_\phi}{2}\eta_t^2 C.
\end{aligned}$$

□

A.12 FedAvg: Bias Under Client Drift (Non-IID)

We formally analyze how multiple local SGD steps in FedAvg introduce *bias* relative to the global DRO gradient $\nabla\phi(\theta)$, and derive the resulting convergence rate. The analysis proceeds in three steps: (i) define the protocol, per-client stochastic gradient, and the effective global direction, (ii) establish conditional unbiasedness and decompose the bias into a heterogeneity term and a client-drift term, and (iii) bound the squared bias and second moment, then telescope the one-step descent to obtain the convergence rate. Throughout this subsection, the local learning rate η is held constant across rounds (i.e., $\eta_t \equiv \eta$ for all t); the optimal choice of η as a function of T is given in Theorem 1.

A.12.1 Protocol and Gradient Definitions

Local update protocol. At round t, a subset $S^{(t)} \subset \mathcal{G}$ of clients participates, each starting from the broadcast model $\theta_{g,0}^{(t)} := \theta^{(t)}$. Throughout the idealized analysis, Assumption 6 specifies fixed-size uniform sampling without replacement, $|S^{(t)}| = m$, fresh local minibatches, and conditional independence of the local sampling randomness across clients. Client g performs K local SGD steps:

$$\theta_{g,k}^{(t)} = \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}, \qquad k = 1, \dots, K, \tag{26}$$

where $\hat{d}_{g,k}^{(t)}$ is the stochastic gradient estimator used at the k-th local step. In this section we analyze the *idealized* setting where each client has access to the exact DRO weights. Concretely, at step k client g samples a minibatch $\{(x_i, \tau_i)\}_{i=1}^{B}$ at the current iterate $\theta_{g,k-1}^{(t)}$ and computes the REINFORCE-based scaled gradient

$$\hat{d}_{g,k}^{(t)} := \xi_g(\theta_{g,k-1}^{(t)}) \cdot \frac{1}{B} \sum_{i=1}^{B} \ell_i\, \nabla_\theta \log \pi_{\theta_{g,k-1}^{(t)}}(\tau_i \mid x_i), \tag{27}$$

where $\xi_g(\theta) = p_g \exp(L_g(\theta)/\lambda) \big/ \sum_h p_h \exp(L_h(\theta)/\lambda)$ is the exact softmax weight based on the true expected losses. The estimator is baseline-free and unclipped; baseline subtraction is allowed only under the conditions in Remark A.4.

Remark A.7 (Relation to Algorithm 2 and 1). Algorithm 2 is the idealized version analyzed in this section: at each local iterate, client g uses the exact weight $\xi_g(\theta)$ computed from all clients' true expected losses $L_h(\theta)$, not from the current minibatch. It performs K local gradient steps with these exact weights. The EMA-extended variant, Algorithm 1, also performs K local gradient steps but uses minibatch costs to update the EMA loss estimator $u_g^{(t)}$ and adds normalization and gradient EMA filters ($\tilde{v}_g$, $\hat{d}_g$), which are updated at each local step. The additional bias introduced by these approximations is analyzed separately in Section A.13 (Stage 2–3, Lemmas A.13–A.16). The present section isolates the effects of heterogeneity and client drift that arise even when the weights are exact.

Server aggregation and effective direction. The server aggregates the client models via FedAvg:

$$\theta^{(t+1)} = \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \theta_{g,K}^{(t)}. \tag{28}$$

Since each client performs K local steps of size η, the aggregation yields

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{|S^{(t)}|} \sum_{g \in S^{(t)}} \sum_{k=1}^{K} \hat{d}_{g,k}^{(t)} = \theta^{(t)} - K\eta\, \hat{d}^{(t)},$$

where we define the *effective global update direction*

$$\hat{d}^{(t)} := \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \hat{d}_{g,k}^{(t)}. \tag{29}$$

The effective global step size is therefore $K\eta$.

A.12.2 Bias Decomposition

Proposition A.1 (Conditionally unbiased local gradients). Under the local gradient definition in Eq. (6) and the regularity conditions of Fact A.1, let $\mathcal{H}_{g,k-1}^{(t)}$ be the local history in Assumption 6, excluding the current minibatch. Conditionally on $\mathcal{H}_{g,k-1}^{(t)}$, draw fresh minibatch inputs $x_i \sim \mathcal{D}_g$ and trajectories $\tau_i \sim \pi_{\theta_{g,k-1}^{(t)}}(\cdot \mid x_i)$. Then, for every participating client g and $k \in \{1, \dots, K\}$,

$$\mathbb{E}\Big[\hat{d}_{g,k}^{(t)} \;\Big|\; \mathcal{H}_{g,k-1}^{(t)}\Big] = \xi_g(\theta_{g,k-1}^{(t)})\nabla L_g(\theta_{g,k-1}^{(t)}) = \nabla\phi_g(\theta_{g,k-1}^{(t)}).$$

Proof. The exact weight $\xi_g(\theta_{g,k-1}^{(t)})$ is a deterministic function of the local iterate and hence is $\mathcal{H}_{g,k-1}^{(t)}$-measurable. Applying Fact A.1 to each fresh sample gives

$$\begin{aligned}\mathbb{E}\Big[\hat{d}_{g,k}^{(t)} \;\Big|\; \mathcal{H}_{g,k-1}^{(t)}\Big] &= \xi_g(\theta_{g,k-1}^{(t)})\frac{1}{B}\sum_{i=1}^{B}\mathbb{E}\Big[\ell_i\nabla_\theta \log \pi_{\theta_{g,k-1}^{(t)}}(\tau_i \mid x_i) \;\Big|\; \mathcal{H}_{g,k-1}^{(t)}\Big] \\ &= \xi_g(\theta_{g,k-1}^{(t)})\nabla L_g(\theta_{g,k-1}^{(t)}) = \nabla\phi_g(\theta_{g,k-1}^{(t)}).\end{aligned}$$

□

Proposition A.2 (Conditional second moment of local gradients). Under Proposition A.1 and Assumptions 3, 5, and 6, for each $g \in S^{(t)}$ and local step j,

$$\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)}\|^2 \;\Big|\; \mathcal{H}_{g,j-1}^{(t)}\Big] \le \sigma^2 + H^2.$$

Proof. For the baseline-free estimator, fresh sampling makes the conditional law of $\hat{d}_{g,j}^{(t)}$ given $\mathcal{H}_{g,j-1}^{(t)}$ depend on that history only through $\theta_{g,j-1}^{(t)}$. Hence Assumption 5 transfers to the finer history:

$$\begin{aligned}&\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)} - \nabla\phi_g(\theta_{g,j-1}^{(t)})\|^2 \;\Big|\; \mathcal{H}_{g,j-1}^{(t)}\Big] \\ &\quad= \mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)} - \nabla\phi_g(\theta_{g,j-1}^{(t)})\|^2 \;\Big|\; \theta_{g,j-1}^{(t)}\Big] \le \sigma^2.\end{aligned}$$

By Proposition A.1, the centered cross term vanishes conditional on $\mathcal{H}_{g,j-1}^{(t)}$. Thus

$$\begin{aligned}\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)}\|^2 \;\Big|\; \mathcal{H}_{g,j-1}^{(t)}\Big] &= \mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)} - \nabla\phi_g(\theta_{g,j-1}^{(t)})\|^2 \;\Big|\; \mathcal{H}_{g,j-1}^{(t)}\Big] \\ &\quad + \|\nabla\phi_g(\theta_{g,j-1}^{(t)})\|^2 \le \sigma^2 + H^2,\end{aligned}$$

where $\|\nabla\phi_g(\theta)\| = \xi_g(\theta)\|\nabla L_g(\theta)\| \le H$. □

Lemma A.8 (FedAvg bias decomposition). Under the local gradient definition in Eq. (6) and Proposition A.1 (conditional unbiasedness of the local stochastic gradient), the effective direction in Eq. (29) satisfies

$$\mathbb{E}\Big[\hat{d}^{(t)} \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] = \frac{1}{G}\nabla\phi(\theta^{(t)}) + b^{(t)},$$

where the bias $b^{(t)}$ decomposes as $b^{(t)} = b_{\text{het}}^{(t)} + b_{\text{drift}}^{(t)}$ with

$$b_{\text{het}}^{(t)} := \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)}), \tag{30}$$

$$b_{\text{drift}}^{(t)} := \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)}) \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big]. \tag{31}$$

Proof. Since $\sigma(\mathcal{F}^{(t)}, S^{(t)}) \subseteq \mathcal{H}_{g,k-1}^{(t)}$, the tower property and Proposition A.1 give

$$\begin{aligned}\mathbb{E}\Big[\hat{d}^{(t)} \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] &= \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\hat{d}_{g,k}^{(t)} \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\ &= \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\mathbb{E}\Big[\hat{d}_{g,k}^{(t)} \;\Big|\; \mathcal{H}_{g,k-1}^{(t)}\Big] \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\ &= \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\nabla\phi_g(\theta_{g,k-1}^{(t)}) \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big], \end{aligned} \tag{32}$$

where the remaining expectation is over local stochastic gradients at steps $1,\dots,k-1$ (which determine $\theta_{g,k-1}^{(t)}$ via Eq. (26)). Add and subtract $\nabla\phi_g(\theta^{(t)})$ inside the sum:

$$\text{Eq. (32)} = \underbrace{\frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\nabla\phi_g(\theta^{(t)})}_{=\frac{1}{G}\nabla\phi(\theta^{(t)})+b_{\text{het}}^{(t)}} + \underbrace{\frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)}) \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big]}_{=b_{\text{drift}}^{(t)}}.$$

□

A.12.3 Bias Bounds and Convergence

Bias and second moment bounds.

Lemma A.9 (Squared bias bound). Under Assumptions 2, 3, 4, 5, and 6, and the local gradient definition in Eq. (6) with Proposition A.1, the squared bias from Lemma A.8 satisfies

$$\mathbb{E}\Big[\|b^{(t)}\|^2 \;\Big|\; \mathcal{F}^{(t)}\Big] \le \frac{2(G-m)}{m(G-1)}\zeta^2 + \eta^2(K-1)^2M_\phi^2\left(\sigma^2+H^2\right), \tag{33}$$

where $M_\phi = M_L + GH^2/(2\lambda)$ is the smoothness constant from Proposition 2. The displayed coefficient is for $G > 1$; for $G = m = 1$ the heterogeneity term is zero and the same drift bound applies.

Proof. By Jensen's inequality (convexity of $\|\cdot\|^2$),

$$\|b^{(t)}\|^2 = \|b_{\text{het}}^{(t)} + b_{\text{drift}}^{(t)}\|^2 \le 2\|b_{\text{het}}^{(t)}\|^2 + 2\|b_{\text{drift}}^{(t)}\|^2.$$

Heterogeneity term. Conditioning on $\mathcal{F}^{(t)}$ fixes $\theta^{(t)}$ and the client gradients. Expanding the squared subset average gives

$$\begin{aligned}\|b_{\text{het}}^{(t)}\|^2 &= \left\|\frac{1}{m}\sum_{g\in S^{(t)}}\left(\nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)})\right)\right\|^2 \\ &= \frac{1}{m^2}\Bigg\{\sum_{g=1}^{G}\mathbf{1}_{\{g\in S^{(t)}\}}\left\|\nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2 \\ &\qquad + \sum_{g\neq h}\mathbf{1}_{\{g\in S^{(t)}\}}\mathbf{1}_{\{h\in S^{(t)}\}}\left\langle\nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)}), \nabla\phi_h(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\rangle\Bigg\}.\end{aligned}$$

Under Assumption 6, for $g \neq h$,

$$\Pr(g\in S^{(t)} \mid \mathcal{F}^{(t)}) = \frac{m}{G}, \qquad \Pr(g,h\in S^{(t)} \mid \mathcal{F}^{(t)}) = \frac{m(m-1)}{G(G-1)}.$$

Since the centered client gradients sum to zero,

$$\sum_{g\neq h}\left\langle \nabla\phi_g(\theta^{(t)})-\frac{1}{G}\nabla\phi(\theta^{(t)}),\nabla\phi_h(\theta^{(t)})-\frac{1}{G}\nabla\phi(\theta^{(t)})\right\rangle$$
$$=-\sum_{g=1}^{G}\left\|\nabla\phi_g(\theta^{(t)})-\frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2.$$

Taking the expectation over $S^{(t)}$ and using Assumption 4,

$$\mathbb{E}\Big[\|b_{\text{het}}^{(t)}\|^2\;\Big|\;\mathcal{F}^{(t)}\Big]=\frac{1}{m^2}\left(\frac{m}{G}-\frac{m(m-1)}{G(G-1)}\right)\sum_{g=1}^{G}\left\|\nabla\phi_g(\theta^{(t)})-\frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2$$
$$=\frac{G-m}{m(G-1)}\frac{1}{G}\sum_{g=1}^{G}\left\|\nabla\phi_g(\theta^{(t)})-\frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2$$
$$\leq\frac{G-m}{m(G-1)}\zeta^2.$$

This is an expectation over the sampled set, not a pointwise bound on every realized subset. For $m=G$, $b_{\text{het}}^{(t)}=0$.

Drift term. We bound $\mathbb{E}[\|b_{\text{drift}}^{(t)}\|^2\mid\mathcal{F}^{(t)}]$ in five steps: (i) express the local iterate drift in terms of accumulated stochastic gradients, (ii) establish that each component map $\nabla\phi_g$ is M_ϕ-Lipschitz, (iii) combine (i) and (ii) to bound the per-step gradient deviation, (iv) bound the second moment of each local stochastic gradient, and (v) average over local steps $k=1,\dots,K$.

Step (i): Local iterate drift. Since $\theta_{g,k-1}^{(t)}$ is obtained from $\theta_{g,0}^{(t)}=\theta^{(t)}$ via $k-1$ local SGD steps using Eq. (26), telescoping gives

$$\theta_{g,k-1}^{(t)}-\theta^{(t)}=\sum_{j=1}^{k-1}(\theta_{g,j}^{(t)}-\theta_{g,j-1}^{(t)})=-\eta\sum_{j=1}^{k-1}\hat{d}_{g,j}^{(t)}. \tag{34}$$

Step (ii): M_ϕ-Lipschitz continuity of $\nabla\phi_g$. We claim that the map $\theta\mapsto\nabla\phi_g(\theta):=\xi_g(\theta)\nabla L_g(\theta)$ is M_ϕ-Lipschitz for each g. By the triangle inequality,

$$\begin{aligned}\|\nabla\phi_g(\theta_1)-\nabla\phi_g(\theta_2)\|&=\|\xi_g(\theta_1)\nabla L_g(\theta_1)-\xi_g(\theta_2)\nabla L_g(\theta_2)\|\\&=\|(\xi_g(\theta_1)-\xi_g(\theta_2))\nabla L_g(\theta_1)+\xi_g(\theta_2)(\nabla L_g(\theta_1)-\nabla L_g(\theta_2))\|\\&\leq\underbrace{|\xi_g(\theta_1)-\xi_g(\theta_2)|\,\|\nabla L_g(\theta_1)\|}_{\text{(I): weight variation}}+\underbrace{\xi_g(\theta_2)\,\|\nabla L_g(\theta_1)-\nabla L_g(\theta_2)\|}_{\text{(II): gradient variation}}.\end{aligned}$$

For (I): by Assumption 3 ($\|\nabla L_g\|\leq H$) and the softmax Lipschitz bound (Lemma A.5, same chain as in the proof of Proposition 2, Term (I)),

$$|\xi_g(\theta_1)-\xi_g(\theta_2)|\leq\|\xi(\theta_1)-\xi(\theta_2)\|_1\leq\frac{1}{2\lambda}\|L(\theta_1)-L(\theta_2)\|_1\leq\frac{GH}{2\lambda}\|\theta_1-\theta_2\|,$$

so (I) $\leq\frac{GH^2}{2\lambda}\|\theta_1-\theta_2\|$.

For (II): since $\xi_g(\theta_2)\leq 1$ and L_g is M_L-smooth (Assumption 2), (II) $\leq M_L\|\theta_1-\theta_2\|$.

Summing (I) and (II):

$$\|\nabla\phi_g(\theta_1)-\nabla\phi_g(\theta_2)\|\leq\underbrace{\left(M_L+\frac{GH^2}{2\lambda}\right)}_{=M_\phi}\|\theta_1-\theta_2\|. \tag{35}$$

Step (iii): Per-step gradient deviation bound. Applying Eq. (35) with $\theta_1=\theta_{g,k-1}^{(t)}$ and $\theta_2=\theta^{(t)}$, and substituting the drift expression in Eq. (34):

$$\|\nabla\phi_g(\theta_{g,k-1}^{(t)})-\nabla\phi_g(\theta^{(t)})\|\leq M_\phi\|\theta_{g,k-1}^{(t)}-\theta^{(t)}\|=M_\phi\,\eta\left\|\sum_{j=1}^{k-1}\hat{d}_{g,j}^{(t)}\right\|. \tag{36}$$

Squaring both sides, taking the conditional expectation given $\mathcal{F}^{(t)}$ and $S^{(t)}$, and applying Jensen's inequality ($\|\sum_{j=1}^{k-1} a_j/(k-1)\|^2 \le \sum_{j=1}^{k-1} \|a_j\|^2/(k-1)$):

$$\mathbb{E}\Big[\|\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)})\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \le M_\phi^2\,\eta^2\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\hat{d}_{g,j}^{(t)}\right\|^2 \;\middle|\; \mathcal{F}^{(t)}, S^{(t)}\right]$$
$$\le M_\phi^2\,\eta^2\,(k{-}1)\sum_{j=1}^{k-1}\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)}\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big]. \quad (37)$$

The Jensen step is for $k \ge 2$; for $k = 1$ the deviation is zero.

Step (iv): Second moment of each local stochastic gradient. By Proposition A.2,

$$\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)}\|^2 \;\Big|\; \mathcal{H}_{g,j-1}^{(t)}\Big] \le \sigma^2 + H^2. \quad (38)$$

Substituting Eq. (38) into Eq. (37) and using the tower property:

$$\begin{aligned}
&\mathbb{E}\Big[\|\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)})\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\
&\le M_\phi^2\eta^2(k-1)\sum_{j=1}^{k-1}\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)}\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\
&= M_\phi^2\eta^2(k-1)\sum_{j=1}^{k-1}\mathbb{E}\Big[\mathbb{E}\Big[\|\hat{d}_{g,j}^{(t)}\|^2 \;\Big|\; \mathcal{H}_{g,j-1}^{(t)}\Big] \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\
&\le M_\phi^2\,\eta^2\,(k{-}1)\sum_{j=1}^{k-1}(\sigma^2 + H^2) \\
&= M_\phi^2\,\eta^2\,(k{-}1)^2\,(\sigma^2 + H^2). \qquad (39)
\end{aligned}$$

Step (v): Averaging over local steps. Since $b_{\text{drift}}^{(t)}$ is an average over $g \in S^{(t)}$ and $k = 1, \dots, K$ (see Eq. (31)), applying Jensen's inequality ($\|\text{average}\|^2 \le$ average of $\|\cdot\|^2$) and substituting Eq. (39) (noting that the $k = 1$ term vanishes):

$$\begin{aligned}
\|b_{\text{drift}}^{(t)}\|^2 &= \left\|\frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)}) \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big]\right\|^2 \\
&\le \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\left\|\mathbb{E}\Big[\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)}) \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big]\right\|^2 \\
&\le \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\|\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)})\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\
&\le \frac{1}{K}\sum_{k=1}^{K} M_\phi^2\,\eta^2\,(k{-}1)^2\,(\sigma^2 + H^2) \\
&= M_\phi^2\,\eta^2\,(\sigma^2 + H^2)\,\frac{(K-1)K(2K-1)}{6K} \\
&\le \frac{1}{2}\eta^2(K-1)^2 M_\phi^2(\sigma^2 + H^2),
\end{aligned}$$

where the last step uses $\frac{2K-1}{6} \le \frac{K-1}{2}$ for $K \ge 2$ (and the bound is trivially 0 for $K = 1$). The same upper bound therefore holds for $\mathbb{E}[\|b_{\text{drift}}^{(t)}\|^2 \mid \mathcal{F}^{(t)}]$. Taking expectations in the initial two-term inequality and combining the heterogeneity and drift bounds gives Eq. (33). □

Lemma A.10 (Second moment bound). Under the same assumptions as Lemma A.9,

$$\mathbb{E}\Big[\|\hat{d}^{(t)}\|^2 \;\Big|\; \mathcal{F}^{(t)}\Big] \le \frac{\sigma^2}{m} + 2H^2 =: C_{\mathrm{fed}}.$$

Proof. By the bias–variance decomposition (the cross term vanishes since $\mathbb{E}[\hat{d}^{(t)} - \mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] \mid \mathcal{F}^{(t)}, S^{(t)}] = 0$ by definition of conditional expectation):

$$\mathbb{E}\Big[\|\hat{d}^{(t)}\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] = \underbrace{\Big\|\mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}]\Big\|^2}_{\text{(I): squared mean}} + \underbrace{\mathbb{E}\Big[\|\hat{d}^{(t)} - \mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}]\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big]}_{\text{(II): variance}}.$$

For (I): from Eq. (32):

$$\mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] = \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\Big[\nabla \phi_g(\theta_{g,k-1}^{(t)}) \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big].$$

Since $\|\nabla \phi_g(\theta)\| = \xi_g(\theta)\|\nabla L_g(\theta)\| \le H$ for every θ (Assumption 3), Jensen's inequality (convexity of $\|\cdot\|$) gives

$$\Big\|\mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}]\Big\| \le \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\Big[\|\nabla \phi_g(\theta_{g,k-1}^{(t)})\| \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \le H,$$

so (I) $\le H^2$.

For (II): write $Y_g := \frac{1}{K}\sum_{k=1}^{K} \hat{d}_{g,k}^{(t)}$ for each client's local average, so that $\hat{d}^{(t)} = \frac{1}{|S^{(t)}|}\sum_{g \in S^{(t)}} Y_g$. By Assumption 6, clients sample independently given $\mathcal{F}^{(t)}$ and $S^{(t)}$, the centered vectors $Y_g - \mathbb{E}[Y_g \mid \mathcal{F}^{(t)}, S^{(t)}]$ ($g \in S^{(t)}$) are mutually independent, so their cross-covariances vanish and

$$\begin{aligned}
\text{(II)} &= \frac{1}{|S^{(t)}|^2}\mathbb{E}\left[\left\|\sum_{g \in S^{(t)}} (Y_g - \mathbb{E}[Y_g \mid \mathcal{F}^{(t)}, S^{(t)}])\right\|^2 \;\middle|\; \mathcal{F}^{(t)}, S^{(t)}\right] \\
&= \frac{1}{|S^{(t)}|^2} \sum_{g \in S^{(t)}} \mathrm{Var}\Big(Y_g \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big). \qquad (40)
\end{aligned}$$

To bound each $\mathrm{Var}(Y_g \mid \mathcal{F}^{(t)}, S^{(t)})$, we use the inequality $\mathrm{Var}(Y_g \mid \mathcal{F}^{(t)}, S^{(t)}) \le \mathbb{E}[\|Y_g\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}]$ and bound the second moment directly. By Jensen's inequality (convexity of $\|\cdot\|^2$):

$$\begin{aligned}
\mathbb{E}[\|Y_g\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}] &= \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K} \hat{d}_{g,k}^{(t)}\right\|^2 \;\middle|\; \mathcal{F}^{(t)}, S^{(t)}\right] \\
&\le \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\hat{d}_{g,k}^{(t)}\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}].
\end{aligned}$$

For each local step k, the tower property and Proposition A.2 give

$$\begin{aligned}
\mathbb{E}[\|\hat{d}_{g,k}^{(t)}\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}] &= \mathbb{E}\Big[\mathbb{E}[\|\hat{d}_{g,k}^{(t)}\|^2 \mid \mathcal{H}_{g,k-1}^{(t)}] \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\
&\le \sigma^2 + H^2.
\end{aligned}$$

Combining: $\mathrm{Var}(Y_g \mid \mathcal{F}^{(t)}, S^{(t)}) \le \sigma^2 + H^2$. Substituting into Eq. (40):

$$\text{(II)} \le \frac{\sigma^2 + H^2}{|S^{(t)}|}.$$

Combining (I) $\le H^2$ and (II) $\le (\sigma^2 + H^2)/|S^{(t)}|$:

$$\mathbb{E}[\|\hat{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}] \le H^2 + \frac{\sigma^2 + H^2}{m} \le \frac{\sigma^2}{m} + 2H^2 = C_{\mathrm{fed}},$$

where the last inequality uses $m \geq 1$. Since the bound does not depend on the realized set $S^{(t)}$,

$$\mathbb{E}[\|\hat{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}] = \mathbb{E}\Big[\mathbb{E}[\|\hat{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}] \;\Big|\; \mathcal{F}^{(t)}\Big] \leq C_{\mathrm{fed}}.$$

□

Convergence to stationary points. Theorem 1 (restated). *Suppose ϕ is M_ϕ-smooth (Proposition 2), Assumptions 3, 4, 5, and 6 hold, and Proposition A.1 applies. With constant step size $\eta > 0$ satisfying $K\eta \leq 1/M_\phi$, the iterates of Algorithm 2 satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\nabla\phi(\theta^{(t)})\|^2\Big] \leq \frac{2G\,\Phi_0}{K\eta T} + \frac{1}{2}G^2\eta^2(K-1)^2M_\phi^2(\sigma^2+H^2) + M_\phi G K\eta\, C_{\mathrm{fed}}, \tag{41}$$

where $\Phi_0 = \phi(\theta^{(0)}) - \phi^\star$, $\phi^\star = \inf_\theta \phi(\theta) > -\infty$, and $C_{\mathrm{fed}} = \sigma^2/m + 2H^2$. All expectations include client selection and local minibatch randomness. For $\Phi_0 > 0$, setting $\eta = \sqrt{2\Phi_0/(M_\phi K^2 T C_{\mathrm{fed}})}$ with $T \geq 2M_\phi\Phi_0/C_{\mathrm{fed}}$ yields the $\mathcal{O}(1/\sqrt{T})$ rate in Eq. (10). The choice $K = 1$ eliminates the drift term and recovers the standard $\mathcal{O}(T^{-1/2})$ rate of centralized SGD.

Proof of Theorem 1. Recall that the global update is $\theta^{(t+1)} = \theta^{(t)} - K\eta\,\hat{d}^{(t)}$ (Section A.12.1). Applying Lemma A.6 with step $K\eta$:

$$\phi(\theta^{(t+1)}) \leq \phi(\theta^{(t)}) - K\eta\langle\nabla\phi(\theta^{(t)}),\, \hat{d}^{(t)}\rangle + \frac{M_\phi}{2}K^2\eta^2\|\hat{d}^{(t)}\|^2.$$

By Lemma A.8 and the tower property,

$$\begin{aligned}
\mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}] &= \mathbb{E}\Big[\mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] \;\Big|\; \mathcal{F}^{(t)}\Big] \\
&= \frac{1}{G}\nabla\phi(\theta^{(t)}) + \mathbb{E}[b^{(t)} \mid \mathcal{F}^{(t)}] \\
&= \frac{1}{G}\nabla\phi(\theta^{(t)}) + \mathbb{E}[b_{\mathrm{het}}^{(t)} \mid \mathcal{F}^{(t)}] + \bar{b}^{(t)},
\end{aligned}$$

where $\bar{b}^{(t)} := \mathbb{E}[b_{\mathrm{drift}}^{(t)} \mid \mathcal{F}^{(t)}]$. Uniform sampling gives

$$\begin{aligned}
\mathbb{E}[b_{\mathrm{het}}^{(t)} \mid \mathcal{F}^{(t)}] &= \mathbb{E}\left[\frac{1}{m}\sum_{g\in S^{(t)}}\nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)}) \;\middle|\; \mathcal{F}^{(t)}\right] \\
&= \frac{1}{m}\sum_{g=1}^{G}\nabla\phi_g(\theta^{(t)})\mathbb{E}[\mathbf{1}_{\{g\in S^{(t)}\}} \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)}) \\
&= \frac{1}{m}\frac{m}{G}\sum_{g=1}^{G}\nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)}) = 0.
\end{aligned}$$

Thus we average over client selection before applying Young's inequality. Taking the conditional expectation of the descent inequality and using Lemma A.10 gives

$$\mathbb{E}[\phi(\theta^{(t+1)}) \mid \mathcal{F}^{(t)}] \leq \phi(\theta^{(t)}) - \frac{K\eta}{G}\|\nabla\phi(\theta^{(t)})\|^2 - K\eta\langle\nabla\phi(\theta^{(t)}),\, \bar{b}^{(t)}\rangle + \frac{M_\phi}{2}K^2\eta^2\, C_{\mathrm{fed}}. \tag{42}$$

For the bias inner product, by Young's inequality ($-\langle a,b\rangle \leq \frac{1}{2\gamma}\|a\|^2 + \frac{\gamma}{2}\|b\|^2$ with $\gamma = G$):

$$-K\eta\langle\nabla\phi(\theta^{(t)}),\, \bar{b}^{(t)}\rangle \leq \frac{K\eta}{2G}\|\nabla\phi(\theta^{(t)})\|^2 + \frac{K\eta G}{2}\|\bar{b}^{(t)}\|^2.$$

Substituting into Eq. (42):

$$\mathbb{E}[\phi(\theta^{(t+1)}) \mid \mathcal{F}^{(t)}] \leq \phi(\theta^{(t)}) - \frac{K\eta}{2G}\|\nabla\phi(\theta^{(t)})\|^2 + \frac{K\eta G}{2}\|\bar{b}^{(t)}\|^2 + \frac{M_\phi}{2}K^2\eta^2\, C_{\mathrm{fed}}.$$

Conditional Jensen's inequality and Step (v) of Lemma A.9 imply

$$\begin{aligned}\|\bar{b}^{(t)}\|^2 &= \|\mathbb{E}[b_{\text{drift}}^{(t)} \mid \mathcal{F}^{(t)}]\|^2 \\ &\le \mathbb{E}[\|b_{\text{drift}}^{(t)}\|^2 \mid \mathcal{F}^{(t)}] \\ &\le \frac{1}{2}\eta^2(K-1)^2 M_\phi^2(\sigma^2+H^2).\end{aligned}$$

Taking full expectation and rearranging,

$$\begin{aligned}&\frac{K\eta}{2G}\,\mathbb{E}\Big[\|\nabla\phi(\theta^{(t)})\|^2\Big] \\ &\le \mathbb{E}[\phi(\theta^{(t)})] - \mathbb{E}[\phi(\theta^{(t+1)})] + \frac{K\eta G}{4}\eta^2(K-1)^2 M_\phi^2(\sigma^2+H^2) + \frac{M_\phi}{2}K^2\eta^2 C_{\text{fed}}.\end{aligned}$$

Summing over $t = 0, \dots, T-1$, and dividing by $K\eta T/(2G)$ yields the telescoping bound in Eq. (41). □

Derivation of the O-notation rate in Eq. (10). *Under the same conditions as Theorem 1, for $\Phi_0 > 0$ set $\eta = \sqrt{2\Phi_0/(M_\phi K^2 T C_{\text{fed}})}$ and assume $T \ge 2M_\phi\Phi_0/C_{\text{fed}}$ (so that $K\eta \le 1/M_\phi$). Then there exists $t^\star \in \{0,\dots,T-1\}$ such that $\mathbb{E}[\|\nabla\phi(\theta^{(t^\star)})\|^2] \le 2G\sqrt{2M_\phi\Phi_0 C_{\text{fed}}/T} + G^2(K-1)^2 M_\phi(\sigma^2+H^2)\Phi_0/(K^2TC_{\text{fed}})$. Since the last term is $\mathcal{O}(1/T)$, this gives $\mathcal{O}(1/\sqrt{T})$. For $K = 1$ the drift term vanishes.*

Proof. Since $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2]$ is bounded by the right-hand side of Eq. (41), there exists $t^\star$ achieving at most this average by the averaging argument. Substituting $\eta = \sqrt{2\Phi_0/(M_\phi K^2 T C_{\text{fed}})}$ into Eq. (41): the first term becomes $G\sqrt{2M_\phi\Phi_0 C_{\text{fed}}/T}$, the last term contributes equally, yielding $2G\sqrt{2M_\phi\Phi_0 C_{\text{fed}}/T}$; the drift coefficient evaluates to $G^2(K-1)^2M_\phi(\sigma^2+H^2)\Phi_0/(K^2TC_{\text{fed}})$. The constraint $K\eta \le 1/M_\phi$ reduces to $T \ge 2M_\phi\Phi_0/C_{\text{fed}}$. □

Discussion. The convergence rate in Eq. (41) has three contributions:

- *Optimization error*: $\mathcal{O}(G/(K\eta T))$ — decreases as $1/\sqrt{T}$;
- *Client drift*: $\mathcal{O}(G^2\eta^2K^2M_\phi^2)$ — grows with the number of local steps, vanishes for $K = 1$, and decreases as $\mathcal{O}(1/T)$ under the stated step size;
- *Stochastic noise*: $\mathcal{O}(M_\phi G K\eta\, C_{\text{fed}})$ — second-moment cost from gradient noise and bounded gradients; decreases at the same rate as the optimization error (since $\eta \propto 1/\sqrt{T}$).

This justifies using $K = 1$ in Algorithm 2 to avoid the client-drift penalty. The EMA-extended variant (Algorithm 1) replaces the exact expected-loss weights with EMA-smoothed quantities; the additional tracking error is analyzed in Section A.13 (see also Remark A.7).

A.12.4 Clipped Local Updates

We separately analyze norm clipping of the exact-DRO-weighted estimator in Eq. (6). All iterates in this subsection follow the clipped recursion below, without momentum or weight decay. The symbols $\tilde{d}_{g,k}^{(t)}$ and $e_{g,k}^{(t)}$ in this subsection denote the clipped direction and clipping error, not EMA tracking quantities. The threshold $H_c > 0$ is distinct from the true-gradient bound H.

Proposition A.3 (Descent bound for clipped local updates). Suppose Assumptions 2, 3, 5, and 6 hold for the iterates below, and the unclipped estimator satisfies the regularity conditions of Fact A.1. Let $\phi^\star = \inf_\theta \phi(\theta) > -\infty$, $\Phi_0 = \phi(\theta^{(0)}) - \phi^\star$, and let the constant step size $\eta > 0$ satisfy $K\eta \le 1/M_\phi$. At each local step, define

$$\tilde{d}_{g,k}^{(t)} = \begin{cases} \hat{d}_{g,k}^{(t)}, & \|\hat{d}_{g,k}^{(t)}\| \le H_c, \\ H_c\hat{d}_{g,k}^{(t)}/\|\hat{d}_{g,k}^{(t)}\|, & \|\hat{d}_{g,k}^{(t)}\| > H_c, \end{cases} \qquad e_{g,k}^{(t)} := \tilde{d}_{g,k}^{(t)} - \hat{d}_{g,k}^{(t)}. \tag{43}$$

Here $\hat{d}_{g,k}^{(t)}$ is the baseline-free estimator in Eq. (6), evaluated at the current clipped local iterate. The local and server updates are

$$\theta_{g,0}^{(t)} = \theta^{(t)}, \qquad \theta_{g,k}^{(t)} = \theta_{g,k-1}^{(t)} - \eta \tilde{d}_{g,k}^{(t)},$$
$$\tilde{d}^{(t)} = \frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \tilde{d}_{g,k}^{(t)}, \qquad \theta^{(t+1)} = \theta^{(t)} - K\eta \tilde{d}^{(t)}.$$

Then, with expectations over client selection and local minibatch randomness,

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla \phi(\theta^{(t)})\|^2 \leq \frac{2G\Phi_0}{K\eta T} + G^2 \eta^2 (K-1)^2 M_\phi^2 (\sigma^2 + H^2) \\ + \frac{2G^2}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} (\|\hat{d}_{g,k}^{(t)}\| - H_c)_+^2 \right] \\ + M_\phi G K \eta (\sigma^2 + H^2). \end{aligned} \tag{44}$$

Here $(\cdot)_+$ denotes the positive part.

Proof. Clipping error and fresh-sample moments. The clipping definition gives

$$\|e_{g,k}^{(t)}\| = (\|\hat{d}_{g,k}^{(t)}\| - H_c)_+, \qquad \|\tilde{d}_{g,k}^{(t)}\| \leq \|\hat{d}_{g,k}^{(t)}\|.$$

Clipping earlier updates changes the current iterate but not the fresh-minibatch conditional law. The proofs of Propositions A.1 and A.2 therefore apply at the clipped iterates and give

$$\mathbb{E}[\hat{d}_{g,k}^{(t)} \mid \mathcal{H}_{g,k-1}^{(t)}] = \nabla \phi_g(\theta_{g,k-1}^{(t)}),$$
$$\mathbb{E}[\|\tilde{d}_{g,k}^{(t)}\|^2 \mid \mathcal{H}_{g,k-1}^{(t)}] \leq \mathbb{E}[\|\hat{d}_{g,k}^{(t)}\|^2 \mid \mathcal{H}_{g,k-1}^{(t)}] \leq \sigma^2 + H^2.$$

Convexity of the norm and of the square also gives

$$\begin{aligned} &\|\mathbb{E}[e_{g,k}^{(t)} \mid \mathcal{H}_{g,k-1}^{(t)}]\|^2 \\ &\quad \leq \left(\mathbb{E}[(\|\hat{d}_{g,k}^{(t)}\| - H_c)_+ \mid \mathcal{H}_{g,k-1}^{(t)}] \right)^2 \\ &\quad \leq \mathbb{E}[(\|\hat{d}_{g,k}^{(t)}\| - H_c)_+^2 \mid \mathcal{H}_{g,k-1}^{(t)}]. \end{aligned}$$

Mean direction after client selection. To distinguish the realized drift along the clipped iterates from $b_{\mathrm{drift}}^{(t)}$ in Lemma A.8, write

$$\begin{aligned} \widetilde{b}_{\mathrm{drift}}^{(t)} &:= \frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \big(\nabla \phi_g(\theta_{g,k-1}^{(t)}) - \nabla \phi_g(\theta^{(t)}) \big), \\ \bar{\widetilde{b}}_{\mathrm{drift}}^{(t)} &:= \mathbb{E}[\widetilde{b}_{\mathrm{drift}}^{(t)} \mid \mathcal{F}^{(t)}], \\ \bar{b}_{\mathrm{clip}}^{(t)} &:= \mathbb{E} \left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} e_{g,k}^{(t)} \,\middle|\, \mathcal{F}^{(t)} \right]. \end{aligned}$$

Using fresh-sample unbiasedness and the tower property,

$$\begin{aligned} \mathbb{E}[\tilde{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] &= \mathbb{E} \left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} (\hat{d}_{g,k}^{(t)} + e_{g,k}^{(t)}) \,\middle|\, \mathcal{F}^{(t)}, S^{(t)} \right] \\ &= \frac{1}{G} \nabla \phi(\theta^{(t)}) + b_{\mathrm{het}}^{(t)} + \mathbb{E}[\widetilde{b}_{\mathrm{drift}}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] \\ &\quad + \mathbb{E} \left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} e_{g,k}^{(t)} \,\middle|\, \mathcal{F}^{(t)}, S^{(t)} \right]. \end{aligned}$$

The term $b_{\mathrm{het}}^{(t)}$ retains the definition in Eq. (30) at the current global iterate. As in the proof of Theorem 1, uniform fixed-size sampling gives $\mathbb{E}[b_{\mathrm{het}}^{(t)} \mid \mathcal{F}^{(t)}] = 0$. Taking the expectation over $S^{(t)}$ before applying Young's inequality,

$$\begin{aligned}\mathbb{E}[\tilde{d}^{(t)} \mid \mathcal{F}^{(t)}] &= \mathbb{E}\Big[\mathbb{E}[\tilde{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] \;\Big|\; \mathcal{F}^{(t)}\Big] \\ &= \frac{1}{G}\nabla\phi(\theta^{(t)}) + \bar{\tilde{b}}_{\mathrm{drift}}^{(t)} + \bar{b}_{\mathrm{clip}}^{(t)}.\end{aligned}$$

Aggregate second moment. Jensen's inequality and the tower property imply

$$\begin{aligned}&\mathbb{E}[\|\tilde{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}] \\ &\quad\le \mathbb{E}\left[\frac{1}{m}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\|\tilde{d}_{g,k}^{(t)}\|^2 \;\middle|\; \mathcal{F}^{(t)}\right] \\ &\quad= \mathbb{E}\left[\frac{1}{m}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\|\tilde{d}_{g,k}^{(t)}\|^2 \mid \mathcal{H}_{g,k-1}^{(t)}] \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \;\middle|\; \mathcal{F}^{(t)}\right] \\ &\quad\le \sigma^2 + H^2.\end{aligned}$$

Drift along the clipped iterates. The local recursion and Cauchy–Schwarz give, for $k \ge 2$,

$$\begin{aligned}&\mathbb{E}[\|\theta_{g,k-1}^{(t)} - \theta^{(t)}\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}] \\ &\quad= \eta^2\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\tilde{d}_{g,j}^{(t)}\right\|^2 \;\middle|\; \mathcal{F}^{(t)}, S^{(t)}\right] \\ &\quad\le \eta^2(k-1)\sum_{j=1}^{k-1}\mathbb{E}[\|\tilde{d}_{g,j}^{(t)}\|^2 \mid \mathcal{F}^{(t)}, S^{(t)}] \\ &\quad= \eta^2(k-1)\sum_{j=1}^{k-1}\mathbb{E}\Big[\mathbb{E}[\|\tilde{d}_{g,j}^{(t)}\|^2 \mid \mathcal{H}_{g,j-1}^{(t)}] \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \\ &\quad\le \eta^2(k-1)^2(\sigma^2 + H^2).\end{aligned}$$

For $k = 1$, the deviation is zero. The component Lipschitz bound in Eq. (35) and Jensen's inequality now yield

$$\begin{aligned}&\|\bar{\tilde{b}}_{\mathrm{drift}}^{(t)}\|^2 \\ &\quad\le \mathbb{E}\left[\frac{1}{m}\sum_{g\in S^{(t)}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\|\nabla\phi_g(\theta_{g,k-1}^{(t)}) - \nabla\phi_g(\theta^{(t)})\|^2 \;\Big|\; \mathcal{F}^{(t)}, S^{(t)}\Big] \;\middle|\; \mathcal{F}^{(t)}\right] \\ &\quad\le \eta^2 M_\phi^2(\sigma^2 + H^2)\frac{1}{K}\sum_{k=1}^{K}(k-1)^2 \\ &\quad\le \frac{1}{2}\eta^2(K-1)^2 M_\phi^2(\sigma^2 + H^2).\end{aligned}$$

The last inequality holds for $K \ge 2$; both sides are zero for $K = 1$.

Clipping bias. Applying Jensen's inequality to the conditional mean and the client/local-step average,

$$\begin{aligned}\|\bar{b}^{(t)}_{\mathrm{clip}}\|^2 &= \left\| \mathbb{E}\left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} e^{(t)}_{g,k} \;\middle|\; \mathcal{F}^{(t)} \right] \right\|^2 \\ &\le \mathbb{E}\left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \|e^{(t)}_{g,k}\|^2 \;\middle|\; \mathcal{F}^{(t)} \right] \\ &= \mathbb{E}\left[\frac{1}{m} \sum_{g \in S^{(t)}} \frac{1}{K} \sum_{k=1}^{K} (\|\hat{d}^{(t)}_{g,k}\| - H_c)^2_+ \;\middle|\; \mathcal{F}^{(t)} \right].\end{aligned}$$

The sampled-set average remains inside the expectation; no pointwise population-dispersion bound on a realized subset is used.

Descent and time averaging. Using the mean direction and second-moment bounds in the descent lemma,

$$\begin{aligned}\mathbb{E}[\phi(\theta^{(t+1)}) \mid \mathcal{F}^{(t)}] \le \phi(\theta^{(t)}) &- \frac{K\eta}{G} \|\nabla\phi(\theta^{(t)})\|^2 \\ &- K\eta \left\langle \nabla\phi(\theta^{(t)}), \bar{\tilde{b}}^{(t)}_{\mathrm{drift}} + \bar{b}^{(t)}_{\mathrm{clip}} \right\rangle + \frac{M_\phi}{2} K^2\eta^2(\sigma^2 + H^2).\end{aligned}$$

Young's inequality bounds the bias inner product by

$$\begin{aligned}&- K\eta \left\langle \nabla\phi(\theta^{(t)}), \bar{\tilde{b}}^{(t)}_{\mathrm{drift}} + \bar{b}^{(t)}_{\mathrm{clip}} \right\rangle \\ &\le \frac{K\eta}{2G} \|\nabla\phi(\theta^{(t)})\|^2 + \frac{K\eta G}{2} \|\bar{\tilde{b}}^{(t)}_{\mathrm{drift}} + \bar{b}^{(t)}_{\mathrm{clip}}\|^2.\end{aligned}$$

Taking full expectations, summing over $t = 0, \dots, T-1$, and dividing by $K\eta T/(2G)$ gives

$$\begin{aligned}\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla\phi(\theta^{(t)})\|^2 \le \frac{2G\Phi_0}{K\eta T} &\\ + \frac{G^2}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\bar{\tilde{b}}^{(t)}_{\mathrm{drift}} + \bar{b}^{(t)}_{\mathrm{clip}}\|^2 &+ M_\phi G K\eta(\sigma^2 + H^2).\end{aligned}$$

Finally, use $\|\bar{\tilde{b}}^{(t)}_{\mathrm{drift}} + \bar{b}^{(t)}_{\mathrm{clip}}\|^2 \le 2\|\bar{\tilde{b}}^{(t)}_{\mathrm{drift}}\|^2 + 2\|\bar{b}^{(t)}_{\mathrm{clip}}\|^2$, substitute the two bias bounds, and use the tower property for the clipping term to obtain Eq. (44). □

Interpretation. For $K = 1$, the drift term vanishes but the clipping term generally remains. This proposition concerns clipping of exact-weight, baseline-free gradients with vanilla SGD, not clipping of the EMA direction in Algorithm 1 or the full training implementation.

Algorithms 3–5 expand the idealized Algorithm 2 and the protocol in Section A.12.1. They assume oracle access to all true expected losses $L_h(\theta)$ at any queried parameter, including each client's current local iterate. The denominator of the exact weight uses all G groups at that same parameter, not only the participating groups or losses at their separate local iterates. This is an analytical reference, not a directly implementable federated baseline. The gradient estimator is baseline-free and unclipped, as in Eq. (6); the conditions for allowing a baseline are stated separately in Remark A.4.

Algorithm 3 One Federated Round of DRO-KL (idealized, exact weights)

Require: Current global model $\theta^{(t)}$, client distributions $\{\mathcal{D}_g\}_{g=1}^{G}$, reference p, temperature $\lambda > 0$, learning rate η, local steps K, batch size B, participation size $m \in \{1, \dots, G\}$, oracle access to all $L_h(\theta)$

Ensure: Updated global model $\theta^{(t+1)}$

Sample $S^{(t)}$ uniformly without replacement with $|S^{(t)}| = m$ under Assumption 6
Server broadcasts $\theta^{(t)}$ to all clients in $S^{(t)}$
for all clients $g \in S^{(t)}$ **in parallel do**
    Run Algorithm 4 with the stated inputs for client g, obtaining $\theta_g^{(t+1)}$
end for
Server collects $\{\theta_g^{(t+1)}\}_{g \in S^{(t)}}$
$\theta^{(t+1)} \leftarrow \textsc{ServerAggregate}(\{\theta_g^{(t+1)}\}_{g \in S^{(t)}})$ ▷ Algorithm 5
return $\theta^{(t+1)}$

Algorithm 4 Local Update for Client g (idealized DRO-KL) — expanded pseudocode

Require: Global model parameters $\theta^{(t)}$, local distribution $\mathcal{D}_g$, reference p, temperature $\lambda > 0$, learning rate η, number of local steps K, batch size B, oracle access to all $L_h(\theta)$

Ensure: Updated local model $\theta_g^{(t+1)}$

$\theta_{g,0}^{(t)} \leftarrow \theta^{(t)}$
for $k = 1, \dots, K$ **do**
    Obtain the exact weight at the current local iterate:

$$\xi_g(\theta_{g,k-1}^{(t)}) = \frac{p_g \exp(L_g(\theta_{g,k-1}^{(t)})/\lambda)}{\sum_{h=1}^{G} p_h \exp(L_h(\theta_{g,k-1}^{(t)})/\lambda)}$$

    Sample a fresh minibatch $x_i \sim \mathcal{D}_g,\ i = 1, \dots, B$
    Sample trajectories $\tau_i \sim \pi_{\theta_{g,k-1}^{(t)}}(\cdot \mid x_i),\ i = 1, \dots, B$
    Compute losses $\ell_i = \ell(\tau_i),\ i = 1, \dots, B$
    Form the exact-weight REINFORCE direction without clipping:

$$\hat{d}_{g,k}^{(t)} \leftarrow \xi_g(\theta_{g,k-1}^{(t)}) \frac{1}{B} \sum_{i=1}^{B} \ell_i \nabla_\theta \log \pi_{\theta_{g,k-1}^{(t)}}(\tau_i \mid x_i)$$

    Local model update:

$$\theta_{g,k}^{(t)} \leftarrow \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}$$

end for
$\theta_g^{(t+1)} \leftarrow \theta_{g,K}^{(t)}$
return $\theta_g^{(t+1)}$

Algorithm 5 Server Aggregation for DRO-KL (idealized)

Require: Client models $\{\theta_g^{(t+1)}\}_{g \in S^{(t)}}$ from the nonempty participating set $S^{(t)}$

Ensure: Updated global model $\theta^{(t+1)}$

Aggregate models (FedAvg):

$$\theta^{(t+1)} \leftarrow \frac{1}{|S^{(t)}|} \sum_{g \in S^{(t)}} \theta_g^{(t+1)}$$

return $\theta^{(t+1)}$

A.13 Convergence Analysis with EMA-Extended Local Updates

In this section we establish the convergence rate of Algorithm 1 (the EMA-extended local update) under full participation ($m = G$) and $K = 1$, following the design in Guo and Yang [2024]. Each client maintains three EMA-smoothed quantities: the loss estimator $u_g^{(t)}$, the normalization constant $\tilde{v}_g$, and the scaled gradient direction $\hat{d}_g$.

A.13.1 Setup: EMA-Extended Local Update

Algorithm description. Let $p_g > 0$, $\sum_g p_g = 1$, $\lambda > 0$, and fix $\beta_u, \beta_v, \beta_d \in (0, 1]$ and $\eta > 0$ within each run. At round t, client g initializes $\theta_{g,0}^{(t)} := \theta^{(t)}$ and performs one local gradient step (Algorithm 1). The client samples a fresh minibatch at $\theta_{g,0}^{(t)}$, updates all three EMA filters (loss u_g, normalization $\tilde{v}_g$, gradient direction $\hat{d}_g$), and takes a model step using the EMA-smoothed gradient direction. For $t \geq 1$, the states satisfy

$$u_{g,0}^{(t)} = u_{g,1}^{(t-1)}, \qquad \tilde{v}_{g,0}^{(t)} = \tilde{v}_{g,1}^{(t-1)}, \qquad \hat{d}_{g,0}^{(t)} = \hat{d}_{g,1}^{(t-1)}.$$

The batch mean loss $\bar{\ell}_g^{(t)}$ and the baseline-free stochastic gradient $\widehat{\nabla L_g}^{(t)}$ may use the same minibatch. Specializing Eqs. (7)–(9) to $K = 1$ gives

$$\begin{aligned}
u_{g,1}^{(t)} &= (1-\beta_u) u_{g,0}^{(t)} + \beta_u \bar{\ell}_g^{(t)}, \\
\hat{w}_{g,1}^{(t)} &= p_g e^{u_{g,1}^{(t)}/\lambda}, \\
\tilde{v}_{g,1}^{(t)} &= (1-\beta_v) \tilde{v}_{g,0}^{(t)} + \beta_v \hat{w}_{g,1}^{(t)}, \\
\hat{d}_{g,1}^{(t)} &= (1-\beta_d) \hat{d}_{g,0}^{(t)} + \beta_d \frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}} \widehat{\nabla L_g}^{(t)}, \\
\theta_{g,1}^{(t)} &= \theta^{(t)} - \eta \hat{d}_{g,1}^{(t)}.
\end{aligned}$$

The server aggregates client models via FedAvg. Since each client performs one local step of size η, the full-participation update is

$$\begin{aligned}
\hat{d}^{(t)} &:= \frac{1}{G} \sum_g \hat{d}_{g,1}^{(t)}, \\
\theta^{(t+1)} &= \theta^{(t)} - \eta \hat{d}^{(t)}, \\
v^{(t+1)} &= \sum_g \tilde{v}_{g,1}^{(t)} = (1-\beta_v) v^{(t)} + \beta_v \sum_g \hat{w}_{g,1}^{(t)}.
\end{aligned} \tag{45}$$

Initialization, history, and input moments. The initial model $\theta^{(0)}$ is fixed. The initial $u_{g,0}^{(0)}$ is an unbiased minibatch estimate of $L_g(\theta^{(0)})$, with $|u_{g,0}^{(0)}| \leq B_L$ almost surely and $\mathbb{E}[|u_{g,0}^{(0)} - L_g(\theta^{(0)})|^2] \leq \sigma_\ell^2$. Initialize

$$\tilde{v}_{g,0}^{(0)} = p_g e^{u_{g,0}^{(0)}/\lambda}, \qquad v^{(0)} = \sum_g \tilde{v}_{g,0}^{(0)}, \qquad \hat{d}_{g,0}^{(0)} = 0.$$

Let $\mathcal{F}^{(t)}$ be generated by the initial states and all minibatch randomness in rounds $s < t$, excluding the current minibatches. Under Assumptions 2, 3, 1, and 7, we require the following fresh-input conditions, almost surely for every g, t:

$$\begin{aligned}
&\mathbb{E}[\bar{\ell}_g^{(t)} \mid \mathcal{F}^{(t)}] = L_g(\theta^{(t)}), \\
&\mathbb{E}[|\bar{\ell}_g^{(t)} - L_g(\theta^{(t)})|^2 \mid \mathcal{F}^{(t)}] \leq \sigma_\ell^2, \\
&\mathbb{E}[\widehat{\nabla L_g}^{(t)} \mid \mathcal{F}^{(t)}] = \nabla L_g(\theta^{(t)}), \\
&\mathbb{E}[\|\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)})\|^2 \mid \mathcal{F}^{(t)}] \leq \sigma^2.
\end{aligned} \tag{46}$$

The gradient mean condition uses the regularity conditions of Fact A.1. We do not assume independence between loss and gradient estimates, independence across clients, almost-sure boundedness of

stochastic gradients, or fourth-moment bounds. In particular, the extra client-independence condition of Assumption 6 is not required here, and Assumption 4 is not used.

Remark A.8 (Server-side normalization in the gradient EMA). In Eq. (9), the gradient direction is scaled by the server-aggregated normalization constant $v^{(t)}$, not the newly computed $v^{(t+1)}$. Thus $u_{g,0}^{(t)}$, $v^{(t)}$, $\theta^{(t)}$, and $\xi_g(\theta^{(t)})$ are $\mathcal{F}^{(t)}$-measurable, but $\hat{w}_{g,1}^{(t)}$ is generally correlated with the current $\widehat{\nabla L_g}^{(t)}$. Although $K = 1$ eliminates within-round local drift, it does not remove the error $v^{(t)} - Z(\theta^{(t)})$ or inter-round tracking delay. These remain in the weight-error part of Lemma A.16.

Notation. We write $Z(\theta) := \sum_{h=1}^{G} p_h \exp(L_h(\theta)/\lambda)$ for the partition function, $w_g(\theta) := p_g \exp(L_g(\theta)/\lambda)$ for the exact (non-EMA) weight (so $\xi_g(\theta) = w_g(\theta)/Z(\theta)$), and $\hat{w}_{g,1}^{(t)} := p_g \exp(u_{g,1}^{(t)}/\lambda)$ for the EMA-based weight. We use $\|X\|_{L^2} := (\mathbb{E}\|X\|^2)^{1/2}$, with unconditional expectations and absolute values for scalar variables. The analysis concerns the uncentered, unclipped updates above, without the implementation's divisive logits, same-batch baseline, BatchNorm, momentum, or weight decay (Appendix A.3).

Lemma A.11 (Bounded cost estimation error). Under Assumption 1 ($|\ell(\tau)| \le B_L$ a.s.), let θ be measurable with respect to a pre-minibatch history $\mathcal{F}$. A fresh minibatch of B conditionally i.i.d. samples $x_i \sim \mathcal{D}_g$, $\tau_i \sim \pi_\theta(\cdot \mid x_i)$ satisfies

$$\mathbb{E}[\bar{\ell}_g(\theta) \mid \mathcal{F}] = L_g(\theta), \qquad \mathbb{E}[|\bar{\ell}_g(\theta) - L_g(\theta)|^2 \mid \mathcal{F}] \le B_L^2/B.$$

Thus $\sigma_\ell^2 = B_L^2/B$ is an admissible bound in Eq. (46).

Proof. Since $\bar{\ell}_g$ is a conditionally unbiased estimator of L_g under the sampling law,

$$\mathbb{E}[|\bar{\ell}_g - L_g|^2 \mid \mathcal{F}] = \mathrm{Var}(\bar{\ell}_g \mid \mathcal{F}) = \frac{1}{B}\mathrm{Var}(\ell(\tau) \mid \mathcal{F}) \le \frac{1}{B}\mathbb{E}[\ell(\tau)^2 \mid \mathcal{F}] \le \frac{B_L^2}{B},$$

using conditional independence within this minibatch and bounded costs. □

Remark A.9 (Scope beyond full participation and $K = 1$). Partial participation introduces retained-state staleness for inactive clients. General K also requires step-level tracking along local iterates and across server aggregation. The round-level recursions proved here do not establish either extension; both remain future work.

A.13.2 Proof Roadmap

Our goal is Theorem 2, which gives a finite-time descent bound for the EMA-extended algorithm. The proof proceeds through four stages, building from a second-moment bound to the final global bound.

Stage 1: Second moment bound (Lemma A.12). We bound $\mathbb{E}[\|\hat{d}^{(t)}\|^2]$ directly from the EMA updates and input conditions, without using the tracking or bias bounds below. This bound also yields the RMS model and loss drift bounds (Proposition A.4) used in Stage 2.

Stage 2: EMA tracking errors (Lemmas A.13–A.15). Each EMA filter introduces a *tracking error* (staleness) because it averages over past iterates where θ was different. We quantify these errors for the loss and normalization EMA filters and an exact-weight comparison gradient EMA. The loss EMA error propagates to the normalization errors (via the exponential map), which in turn affect the scaled gradient EMA.

Stage 3: Bias decomposition (Lemma A.16). We show $\mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}] = G^{-1}\nabla\phi(\theta^{(t)}) + b^{(t)}$, where the total bias $b^{(t)}$ is decomposed into gradient staleness and weight-estimation error.

Stage 4: Convergence rate (Theorem 2). Combining the M_ϕ-smoothness descent lemma (Lemma A.6) with the bias (Stage 3) and moment bounds (Stage 1), we telescope over T rounds to obtain the finite-time bound.

A.13.3 Stage 1: Second Moment Bound

All results below use the full-participation, single-step setting, initialization, and input conditions of Section A.13.1.

Lemma A.12 (Second moment with EMA). Consider Algorithm 1 with $m = G$, $K = 1$, $p_g > 0$, $\sum_g p_g = 1$, $\lambda > 0$, and fixed $\beta_u, \beta_v, \beta_d \in (0, 1]$. Assume $|u_{g,0}^{(0)}| \le B_L$, $\tilde{v}_{g,0}^{(0)} = p_g e^{u_{g,0}^{(0)}/\lambda}$, $v^{(0)} = \sum_g \tilde{v}_{g,0}^{(0)}$, and $\hat{d}_{g,0}^{(0)} = 0$, with all states retained between rounds. Under Assumptions 1, 3, and 7, suppose the fresh gradient input is conditionally unbiased given the pre-minibatch history $\mathcal{F}^{(t)}$. Then the actual aggregate direction satisfies, for every $t \ge 0$,

$$\mathbb{E}[\|\hat{d}^{(t)}\|^2] \le \frac{e^{4B_L/\lambda}}{G^2}(H^2 + \sigma^2) =: D_{\mathrm{ema}}^2, \qquad \hat{d}^{(t)} := \frac{1}{G}\sum_g \hat{d}_{g,1}^{(t)}. \tag{47}$$

The expectation is unconditional. No independence between the loss and gradient inputs, or between clients, is required.

Proof. Convexity of the EMA updates and bounded costs imply, almost surely,

$$|u_{g,1}^{(t)}| \le B_L, \qquad p_g e^{-B_L/\lambda} \le \tilde{v}_{g,1}^{(t)} \le p_g e^{B_L/\lambda}, \qquad e^{-B_L/\lambda} \le v^{(t)} \le e^{B_L/\lambda}.$$

In particular, the old denominator used by the actual update gives $0 < \hat{w}_{g,1}^{(t)}/v^{(t)} \le p_g e^{2B_L/\lambda}$. By the bias–variance decomposition, conditional unbiasedness and Assumptions 3 and 7 give

$$\begin{aligned}\mathbb{E}[\|\widehat{\nabla L_g}^{(t)}\|^2 \mid \mathcal{F}^{(t)}] &= \|\nabla L_g(\theta^{(t)})\|^2 + \mathbb{E}[\|\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)})\|^2 \mid \mathcal{F}^{(t)}] \\ &\le H^2 + \sigma^2.\end{aligned}$$

Write $\|X\|_{L^2} := (\mathbb{E}\|X\|^2)^{1/2}$. Using the coefficient bound inside the product and then taking full expectations,

$$\left\| \frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}} \widehat{\nabla L_g}^{(t)} \right\|_{L^2} \le p_g e^{2B_L/\lambda}\sqrt{H^2 + \sigma^2}.$$

Zero initialization and state retention give the exact expansion

$$\hat{d}_{g,1}^{(t)} = \beta_d \sum_{s=0}^{t} (1-\beta_d)^{t-s} \frac{\hat{w}_{g,1}^{(s)}}{v^{(s)}} \widehat{\nabla L_g}^{(s)}.$$

Minkowski's inequality therefore yields

$$\begin{aligned}\|\hat{d}_{g,1}^{(t)}\|_{L^2} &\le \beta_d \sum_{s=0}^{t} (1-\beta_d)^{t-s} \left\| \frac{\hat{w}_{g,1}^{(s)}}{v^{(s)}} \widehat{\nabla L_g}^{(s)} \right\|_{L^2} \\ &\le p_g e^{2B_L/\lambda}\sqrt{H^2 + \sigma^2}\left[1 - (1-\beta_d)^{t+1}\right].\end{aligned}$$

Applying Minkowski's inequality once more and using $\sum_g p_g = 1$,

$$\begin{aligned}\|\hat{d}^{(t)}\|_{L^2} &\le \frac{1}{G}\sum_g \|\hat{d}_{g,1}^{(t)}\|_{L^2} \\ &\le \frac{e^{2B_L/\lambda}}{G}\sqrt{H^2 + \sigma^2}\left[1 - (1-\beta_d)^{t+1}\right] \le \frac{e^{2B_L/\lambda}}{G}\sqrt{H^2 + \sigma^2}.\end{aligned}$$

Squaring proves the claim. □

Comparison with the non-EMA case. When $\beta_u = \beta_v = \beta_d = 1$, the current loss estimate still contains sampling noise, and the gradient update still uses the previous round's normalization $v^{(t)}$. Thus the actual input need not equal the exact-weight estimator, even though the direction EMA has no memory.

A.13.4 Stage 2: EMA Tracking Errors

Proposition A.4 (RMS model and loss drift). In the setting of Section A.13.1, define

$$\rho_{\theta,\mathrm{RMS}} := \frac{\eta e^{2B_L/\lambda}}{G}\sqrt{H^2+\sigma^2}.$$

For every g, t,

$$\|\theta^{(t+1)}-\theta^{(t)}\|_{L^2} \le \rho_{\theta,\mathrm{RMS}}, \qquad \|L_g(\theta^{(t+1)})-L_g(\theta^{(t)})\|_{L^2} \le H\rho_{\theta,\mathrm{RMS}}.$$

Proof. Since $\theta^{(t+1)} = \theta^{(t)} - \eta\hat{d}^{(t)}$, Lemma A.12 and Eq. (45) give

$$\|\theta^{(t+1)}-\theta^{(t)}\|_{L^2} = \eta\|\hat{d}^{(t)}\|_{L^2} \le \frac{\eta e^{2B_L/\lambda}}{G}\sqrt{H^2+\sigma^2}.$$

The mean value theorem and $\|\nabla L_g\| \le H$ imply $|L_g(\theta') - L_g(\theta)| \le H\|\theta'-\theta\|$. Taking the L^2 norm gives the loss-drift bound. □

2a. Loss EMA tracking error. We first quantify how the EMA loss estimator $u_{g,1}^{(t)}$ tracks the true loss $L_g(\theta^{(t)})$.

Lemma A.13 (Loss EMA tracking error). In the setting of Section A.13.1, for every g, t,

$$\mathbb{E}[|u_{g,1}^{(t)} - L_g(\theta^{(t)})|^2] \le \max\left\{\sigma_\ell^2,\ \beta_u\sigma_\ell^2 + \frac{(1-\beta_u)^2}{\beta_u^2}H^2\rho_{\theta,\mathrm{RMS}}^2\right\} =: \epsilon_u^2.$$

Proof. For $t \ge 1$, the retained-state recursion gives

$$\begin{aligned} u_{g,1}^{(t)} - L_g(\theta^{(t)}) = (1-\beta_u)\big[u_{g,1}^{(t-1)} - L_g(\theta^{(t-1)}) - (L_g(\theta^{(t)}) - L_g(\theta^{(t-1)}))\big] \\ + \beta_u(\bar{\ell}_g^{(t)} - L_g(\theta^{(t)})). \end{aligned}$$

The bracket is $\mathcal{F}^{(t)}$-measurable. Squaring and conditioning on $\mathcal{F}^{(t)}$ removes its cross term with the current loss noise by Eq. (46), giving

$$\begin{aligned} &\mathbb{E}[|u_{g,1}^{(t)} - L_g(\theta^{(t)})|^2 \mid \mathcal{F}^{(t)}] \\ &\quad= (1-\beta_u)^2|u_{g,1}^{(t-1)} - L_g(\theta^{(t-1)}) - (L_g(\theta^{(t)}) - L_g(\theta^{(t-1)}))|^2 \\ &\qquad+ \beta_u^2\mathbb{E}[|\bar{\ell}_g^{(t)} - L_g(\theta^{(t)})|^2 \mid \mathcal{F}^{(t)}]. \end{aligned}$$

To handle the cross term between the previous tracking residual and the loss drift, we apply the parametric Young's inequality with $\gamma = \beta_u/(1-\beta_u)$ for $0 < \beta_u < 1$. Taking full expectations and using Proposition A.4 yields

$$\begin{aligned} \mathbb{E}[|u_{g,1}^{(t)} - L_g(\theta^{(t)})|^2] &\le (1-\beta_u)\mathbb{E}[|u_{g,1}^{(t-1)} - L_g(\theta^{(t-1)})|^2] \\ &\quad+ \frac{(1-\beta_u)^2}{\beta_u}\mathbb{E}[|L_g(\theta^{(t)}) - L_g(\theta^{(t-1)})|^2] + \beta_u^2\sigma_\ell^2 \\ &\le (1-\beta_u)\mathbb{E}[|u_{g,1}^{(t-1)} - L_g(\theta^{(t-1)})|^2] \\ &\quad+ \frac{(1-\beta_u)^2}{\beta_u}H^2\rho_{\theta,\mathrm{RMS}}^2 + \beta_u^2\sigma_\ell^2. \end{aligned}$$

At $t = 0$, convexity of the square, the initial error bound, and the current loss variance bound give $\mathbb{E}[|u_{g,1}^{(0)} - L_g(\theta^{(0)})|^2] \le \sigma_\ell^2$. Since $(1-\beta_u) < 1$, unrolling the recursion from this initial post-update error yields, for $t \ge 1$,

$$\begin{aligned} \mathbb{E}[|u_{g,1}^{(t)} - L_g(\theta^{(t)})|^2] &\le (1-\beta_u)^t\sigma_\ell^2 \\ &\quad+ \left[1-(1-\beta_u)^t\right]\left(\beta_u\sigma_\ell^2 + \frac{(1-\beta_u)^2}{\beta_u^2}H^2\rho_{\theta,\mathrm{RMS}}^2\right) \le \epsilon_u^2. \end{aligned}$$

For $\beta_u = 1$, the estimator is the current batch mean and the same bound follows directly from Eq. (46). □

2b. Normalization constant EMA tracking error. The loss EMA tracking error ϵ_u propagates to the weight via the exponential map. The normalization EMA then incurs additional staleness. We track the actual old denominator against $Z(\theta^{(t)})$, including inter-round model movement.

Lemma A.14 (Normalization EMA tracking error). In the setting of Section A.13.1,

$$\|\hat{w}_{g,1}^{(t)} - w_g(\theta^{(t)})\|_{L^2} \le \frac{p_g e^{B_L/\lambda}}{\lambda}\epsilon_u, \qquad \|v^{(t)} - Z(\theta^{(t)})\|_{L^2} \le V_t,$$

where the deterministic bound is

$$V_t := \frac{e^{B_L/\lambda}}{\lambda}\left[(1-\beta_v)^t\sigma_\ell + \left(1-(1-\beta_v)^t\right)\left(\epsilon_u + \frac{H\rho_{\theta,\mathrm{RMS}}}{\beta_v}\right)\right].$$

We use $(1-\beta_v)^0 = 1$, also when $\beta_v = 1$.

Proof. Convexity of the EMA and bounded costs keep all loss states in $[-B_L, B_L]$. The exponential map $f(x) = e^{x/\lambda}$ has derivative bounded by $e^{B_L/\lambda}/\lambda$ on this interval. By the mean value theorem,

$$|\hat{w}_{g,1}^{(t)} - w_g(\theta^{(t)})| \le \frac{p_g e^{B_L/\lambda}}{\lambda}|u_{g,1}^{(t)} - L_g(\theta^{(t)})|.$$

Taking L^2 norms and applying Lemma A.13 proves the weight estimate. Moreover, $\|\nabla Z(\theta)\| \le \sum_g p_g e^{B_L/\lambda} H/\lambda = He^{B_L/\lambda}/\lambda$. The server recursion therefore gives

$$\begin{aligned} v^{(t+1)} - Z(\theta^{(t+1)}) &= (1-\beta_v)(v^{(t)} - Z(\theta^{(t)})) \\ &\quad + \beta_v \sum_g (\hat{w}_{g,1}^{(t)} - w_g(\theta^{(t)})) + Z(\theta^{(t)}) - Z(\theta^{(t+1)}). \end{aligned}$$

Minkowski's inequality, $\sum_g p_g = 1$, and the RMS movement bound give

$$\begin{aligned} \|v^{(t+1)} - Z(\theta^{(t+1)})\|_{L^2} &\le (1-\beta_v)\|v^{(t)} - Z(\theta^{(t)})\|_{L^2} \\ &\quad + \frac{e^{B_L/\lambda}}{\lambda}(\beta_v\epsilon_u + H\rho_{\theta,\mathrm{RMS}}). \end{aligned}$$

The initialization satisfies

$$\begin{aligned} \|v^{(0)} - Z(\theta^{(0)})\|_{L^2} &\le \frac{e^{B_L/\lambda}}{\lambda}\sum_g p_g \|u_{g,0}^{(0)} - L_g(\theta^{(0)})\|_{L^2} \\ &\le \frac{e^{B_L/\lambda}}{\lambda}\sigma_\ell. \end{aligned}$$

Unrolling the scalar norm recursion yields V_t, including $t = 0$. No independence is needed in these triangle inequalities. □

2c. Gradient direction EMA tracking error. To separate the EMA-intrinsic effects from the weight-approximation error, we analyze a comparison gradient EMA whose input at each round uses the exact weight and the same stochastic gradient at the actual iterate. The discrepancy from the actual direction is bounded in Stage 3 (Lemma A.16).

Lemma A.15 (Tracking of the exact-weight comparison EMA). We define $b_{\mathrm{stale}}^{(t)}$ for the bias decomposition in Stage 3 (Lemma A.16). In the setting of Section A.13.1, define

$$\begin{aligned} \hat{d}_{g,1}^{(t),\mathrm{ideal}} &= (1-\beta_d)\hat{d}_{g,0}^{(t),\mathrm{ideal}} + \beta_d \xi_g(\theta^{(t)})\widehat{\nabla L_g}^{(t)}, \\ \hat{d}_{g,0}^{(0),\mathrm{ideal}} &= 0, \\ \hat{d}_{g,0}^{(t),\mathrm{ideal}} &= \hat{d}_{g,1}^{(t-1),\mathrm{ideal}} \quad (t \ge 1), \\ \hat{d}^{(-1),\mathrm{ideal}} &:= 0, \\ \hat{d}^{(t),\mathrm{ideal}} &:= \frac{1}{G}\sum_g \hat{d}_{g,1}^{(t),\mathrm{ideal}}, \\ b_{\mathrm{stale}}^{(t)} &:= \mathbb{E}[\hat{d}^{(t),\mathrm{ideal}} \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)}). \end{aligned}$$

The deterministic nonnegative sequences

$$
\begin{aligned}
R_0 &:= \frac{(1-\beta_d)^2 H^2 + \beta_d^2 \sigma^2}{G^2}, \\
B_0 &:= \frac{(1-\beta_d)^2 H^2}{G^2}, \\
R_t &:= (1-\beta_d) R_{t-1} + \frac{(1-\beta_d)^2 M_\phi^2 \rho_{\theta,\mathrm{RMS}}^2}{\beta_d G^2} + \frac{\beta_d^2 \sigma^2}{G^2}, && t \geq 1, \\
B_t &:= (1-\beta_d) R_{t-1} + \frac{(1-\beta_d)^2 M_\phi^2 \rho_{\theta,\mathrm{RMS}}^2}{\beta_d G^2}, && t \geq 1
\end{aligned}
$$

satisfy

$$
\begin{aligned}
&\mathbb{E}[\|\hat{d}^{(t),\mathrm{ideal}} - G^{-1}\nabla\phi(\theta^{(t)})\|^2] \leq R_t, \\
&\mathbb{E}[\|b_{\mathrm{stale}}^{(t)}\|^2] \leq B_t.
\end{aligned}
$$

Proof. The previous comparison state is $\mathcal{F}^{(t)}$-measurable. Conditional unbiasedness of only the current input gives

$$
\begin{aligned}
\mathbb{E}[\hat{d}^{(t),\mathrm{ideal}} \mid \mathcal{F}^{(t)}] &= (1-\beta_d)\hat{d}^{(t-1),\mathrm{ideal}} + \frac{\beta_d}{G}\nabla\phi(\theta^{(t)}), \\
b_{\mathrm{stale}}^{(t)} &= (1-\beta_d)\left(\hat{d}^{(t-1),\mathrm{ideal}} - \frac{1}{G}\nabla\phi(\theta^{(t)})\right).
\end{aligned}
$$

Since the exact weights are history-measurable, nonnegative, and sum to one, conditional Jensen's inequality yields

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{G}\sum_g \xi_g(\theta^{(t)})(\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)}))\right\|^2 \,\middle|\, \mathcal{F}^{(t)}\right] \\
&\quad\leq \frac{1}{G^2}\sum_g \xi_g(\theta^{(t)})\mathbb{E}[\|\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)})\|^2 \mid \mathcal{F}^{(t)}] \leq \frac{\sigma^2}{G^2}.
\end{aligned}
$$

This noise has conditional mean zero, without client independence. Writing $r^{(t)} := \hat{d}^{(t),\mathrm{ideal}} - G^{-1}\nabla\phi(\theta^{(t)})$, we have

$$
r^{(t)} = b_{\mathrm{stale}}^{(t)} + \frac{\beta_d}{G}\sum_g \xi_g(\theta^{(t)})(\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)})).
$$

Squaring, conditioning on $\mathcal{F}^{(t)}$, and then taking full expectations removes the cross term and gives

$$
\mathbb{E}[\|r^{(t)}\|^2] \leq \mathbb{E}[\|b_{\mathrm{stale}}^{(t)}\|^2] + \frac{\beta_d^2\sigma^2}{G^2}.
$$

At $t=0$, $b_{\mathrm{stale}}^{(0)} = -(1-\beta_d)\nabla\phi(\theta^{(0)})/G$; hence the two initial bounds are B_0 and R_0, respectively. For $t \geq 1$, Young's inequality gives

$$
\begin{aligned}
\|b_{\mathrm{stale}}^{(t)}\|^2 &= (1-\beta_d)^2\left\|r^{(t-1)} - \frac{\nabla\phi(\theta^{(t)}) - \nabla\phi(\theta^{(t-1)})}{G}\right\|^2 \\
&\leq (1-\beta_d)\|r^{(t-1)}\|^2 + \frac{(1-\beta_d)^2}{\beta_d G^2}\|\nabla\phi(\theta^{(t)}) - \nabla\phi(\theta^{(t-1)})\|^2.
\end{aligned}
$$

Using M_ϕ-smoothness and Proposition A.4,

$$
\begin{aligned}
\mathbb{E}[\|b_{\mathrm{stale}}^{(t)}\|^2] &\leq (1-\beta_d)\mathbb{E}[\|r^{(t-1)}\|^2] + \frac{(1-\beta_d)^2 M_\phi^2 \rho_{\theta,\mathrm{RMS}}^2}{\beta_d G^2}, \\
\mathbb{E}[\|r^{(t)}\|^2] &\leq (1-\beta_d)\mathbb{E}[\|r^{(t-1)}\|^2] + \frac{(1-\beta_d)^2 M_\phi^2 \rho_{\theta,\mathrm{RMS}}^2}{\beta_d G^2} + \frac{\beta_d^2\sigma^2}{G^2}.
\end{aligned}
$$

Induction gives the claimed bounds. For $\beta_d = 1$, the conditional staleness bias is zero and the same inequalities hold directly. □

A.13.5 Stage 3: Bias Decomposition

We now combine the tracking errors from Stage 2 to characterize the total bias of $\hat{d}^{(t)}$. With $m = G$ and $K = 1$, $\sum_g \nabla\phi_g = \nabla\phi$ and $\theta_{g,0}^{(t)} = \theta^{(t)}$ give

$$b_{\text{het}}^{(t)} = \frac{1}{G}\sum_g \nabla\phi_g(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t)}) = 0, \tag{48}$$

$$b_{\text{drift}}^{(t)} = 0. \tag{49}$$

Lemma A.16 (Bias decomposition with EMA). In the setting of Section A.13.1, let

$$A_t := \frac{e^{2B_L/\lambda}}{\lambda}\epsilon_u + e^{3B_L/\lambda}V_t,$$
$$b^{(t)} := \mathbb{E}[\hat{d}^{(t)} \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)}),$$
$$b_{\text{wt}}^{(t)} := \mathbb{E}[\hat{d}^{(t)} - \hat{d}^{(t),\text{ideal}} \mid \mathcal{F}^{(t)}].$$

Then $b^{(t)} = b_{\text{stale}}^{(t)} + b_{\text{wt}}^{(t)}$, with $b_{\text{stale}}^{(t)}$ as in Lemma A.15, and

$$\mathbb{E}[\|b_{\text{stale}}^{(t)}\|^2] \le B_t, \tag{50}$$

$$\|b_{\text{wt}}^{(t)}\|_{L^2} \le \frac{\beta_d}{G}\sum_{s=0}^{t}(1-\beta_d)^{t-s}(HA_s + e^{2B_L/\lambda}\sigma). \tag{51}$$

In particular, for every $t \ge 0$,

$$\mathbb{E}[\|b^{(t)}\|^2] \le \left[\sqrt{B_t} + \frac{\beta_d}{G}\sum_{s=0}^{t}(1-\beta_d)^{t-s}(HA_s + e^{2B_L/\lambda}\sigma)\right]^2 =: Q_t. \tag{52}$$

All outer expectations and L^2 norms are unconditional, and $(1-\beta_d)^0 = 1$, including when $\beta_d = 1$.

Proof. **Comparison and bias decomposition.** Linearity and $\sum_g \nabla\phi_g = \nabla\phi$ give

$$b_{\text{stale}}^{(t)} = \frac{1}{G}\sum_g \mathbb{E}[\hat{d}_{g,1}^{(t),\text{ideal}} - \nabla\phi_g(\theta^{(t)}) \mid \mathcal{F}^{(t)}],$$
$$b^{(t)} = b_{\text{stale}}^{(t)} + b_{\text{wt}}^{(t)}.$$

With $\hat{d}^{(-1)} = \hat{d}^{(-1),\text{ideal}} = 0$, the actual and comparison updates satisfy the pathwise identity

$$\hat{d}^{(t)} - \hat{d}^{(t),\text{ideal}} = (1-\beta_d)(\hat{d}^{(t-1)} - \hat{d}^{(t-1),\text{ideal}}) + \frac{\beta_d}{G}\sum_g \left(\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}} - \xi_g(\theta^{(t)})\right)\widehat{\nabla L_g}^{(t)}.$$

The stochastic gradient in the last product cannot be replaced by its mean while leaving the current random weight unchanged.

Coefficient error. Define, as a difference of the actual and exact coefficients,

$$c_g^{(t)} := \frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}} - \xi_g(\theta^{(t)}).$$

Bounded loss states and initialization give $e^{-B_L/\lambda} \le v^{(t)}, Z(\theta^{(t)}) \le e^{B_L/\lambda}$ almost surely. The identity

$$c_g^{(t)} = \frac{\hat{w}_{g,1}^{(t)} - w_g(\theta^{(t)})}{v^{(t)}} + w_g(\theta^{(t)})\left(\frac{1}{v^{(t)}} - \frac{1}{Z(\theta^{(t)})}\right)$$

and Lemma A.14 imply

$$\begin{aligned}\|c_g^{(t)}\|_{L^2} &\le e^{B_L/\lambda}\|\hat{w}_{g,1}^{(t)} - w_g(\theta^{(t)})\|_{L^2} + p_g e^{3B_L/\lambda}\|v^{(t)} - Z(\theta^{(t)})\|_{L^2} \\ &\le p_g\left(\frac{e^{2B_L/\lambda}}{\lambda}\epsilon_u + e^{3B_L/\lambda}V_t\right) = p_g A_t.\end{aligned}$$

Both $\hat{w}_{g,1}^{(t)}/v^{(t)}$ and $\xi_g(\theta^{(t)})$ lie in $[0, p_g e^{2B_L/\lambda}]$; hence $|c_g^{(t)}| \le p_g e^{2B_L/\lambda}$ almost surely.

Product and accumulated error. Splitting the true-gradient and noise parts, and using this last bound inside the noise product, gives

$$\begin{aligned}\|c_g^{(t)}\widehat{\nabla L_g}^{(t)}\|_{L^2} &\le H\|c_g^{(t)}\|_{L^2} + \|c_g^{(t)}(\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)}))\|_{L^2} \\ &\le p_g H A_t + p_g e^{2B_L/\lambda}\|\widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)})\|_{L^2} \\ &\le p_g(HA_t + e^{2B_L/\lambda}\sigma).\end{aligned}$$

Unrolling the difference recursion and applying Minkowski's inequality,

$$\begin{aligned}\|\hat{d}^{(t)} - \hat{d}^{(t),\mathrm{ideal}}\|_{L^2} &\le \frac{\beta_d}{G}\sum_{s=0}^{t}(1-\beta_d)^{t-s}\sum_g \|c_g^{(s)}\widehat{\nabla L_g}^{(s)}\|_{L^2} \\ &\le \frac{\beta_d}{G}\sum_{s=0}^{t}(1-\beta_d)^{t-s}(HA_s + e^{2B_L/\lambda}\sigma).\end{aligned}$$

Conditional Jensen's inequality gives $\|b_{\mathrm{wt}}^{(t)}\|_{L^2} \le \|\hat{d}^{(t)} - \hat{d}^{(t),\mathrm{ideal}}\|_{L^2}$. Combining with Lemma A.15 and another triangle inequality yields

$$\|b^{(t)}\|_{L^2} \le \sqrt{B_t} + \frac{\beta_d}{G}\sum_{s=0}^{t}(1-\beta_d)^{t-s}(HA_s + e^{2B_L/\lambda}\sigma).$$

Squaring proves Eq. (52). □

A.13.6 Stage 4: Convergence Rate

We first state the one-step progress bound (analogous to Lemma A.7), then telescope to obtain the final bound. We condition on the pre-minibatch history, then take full expectations before applying the unconditional moment bounds.

Lemma A.17 (One-step progress for the EMA update). In the setting of Section A.13.1, with the conditional bias from Lemma A.16,

$$\begin{aligned}\mathbb{E}[\phi(\theta^{(t+1)}) \mid \mathcal{F}^{(t)}] \le \phi(\theta^{(t)}) - \frac{\eta}{2G}\|\nabla\phi(\theta^{(t)})\|^2 \\ + \frac{\eta G}{2}\|b^{(t)}\|^2 + \frac{M_\phi\eta^2}{2}\mathbb{E}[\|\hat{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}].\end{aligned} \tag{53}$$

Proof. The descent lemma (Lemma A.6) holds pathwise:

$$\phi(\theta^{(t+1)}) \le \phi(\theta^{(t)}) - \eta\langle\nabla\phi(\theta^{(t)}), \hat{d}^{(t)}\rangle + \frac{M_\phi\eta^2}{2}\|\hat{d}^{(t)}\|^2.$$

Since $\theta^{(t)}$ and $b^{(t)}$ are $\mathcal{F}^{(t)}$-measurable,

$$\begin{aligned}\mathbb{E}[\phi(\theta^{(t+1)}) \mid \mathcal{F}^{(t)}] \le \phi(\theta^{(t)}) - \frac{\eta}{G}\|\nabla\phi(\theta^{(t)})\|^2 - \eta\langle\nabla\phi(\theta^{(t)}), b^{(t)}\rangle \\ + \frac{M_\phi\eta^2}{2}\mathbb{E}[\|\hat{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}].\end{aligned}$$

For the bias inner product, Young's inequality gives

$$-\eta\langle\nabla\phi(\theta^{(t)}), b^{(t)}\rangle \le \frac{\eta}{2G}\|\nabla\phi(\theta^{(t)})\|^2 + \frac{\eta G}{2}\|b^{(t)}\|^2.$$

Substitution proves Eq. (53). □

Theorem 2 (restated, explicit form). *Use the setting, initialization, and conditional input moments of Section A.13.1, with $m = G$, $K = 1$, and fixed $\eta > 0$, $\beta_u, \beta_v, \beta_d \in (0, 1]$ within each run. Let $\phi^\star = \inf_\theta \phi(\theta) > -\infty$ and $\Phi_0 = \phi(\theta^{(0)}) - \phi^\star$. For every integer $T \geq 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] \leq \frac{2G\Phi_0}{\eta T} + \frac{G^2}{T}\sum_{t=0}^{T-1} Q_t + M_\phi G\eta D_{\mathrm{ema}}^2. \tag{54}$$

The minimum over $0 \leq t < T$ of the expected squared gradient norm obeys the same bound. Here Q_t is given by Eq. (52) in Lemma A.16 and $D_{\mathrm{ema}}^2 = e^{4B_L/\lambda}(H^2 + \sigma^2)/G^2$ in Lemma A.12. An attained minimum of ϕ is not assumed. Bounded losses imply $-B_L \leq \phi(\theta) \leq B_L$, ensuring the lower bound and finiteness of the expectations.

Proof of Theorem 2. Taking full expectations in Lemma A.17 first, and only then applying Lemmas A.16 and A.12, gives

$$\begin{aligned}\mathbb{E}[\phi(\theta^{(t+1)})] &\leq \mathbb{E}[\phi(\theta^{(t)})] - \frac{\eta}{2G}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] \\ &\quad + \frac{\eta G}{2}\mathbb{E}[\|b^{(t)}\|^2] + \frac{M_\phi\eta^2}{2}\mathbb{E}[\|\hat{d}^{(t)}\|^2] \\ &\leq \mathbb{E}[\phi(\theta^{(t)})] - \frac{\eta}{2G}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] + \frac{\eta G}{2}Q_t + \frac{M_\phi\eta^2}{2}D_{\mathrm{ema}}^2.\end{aligned}$$

Rearranging and summing over $t = 0, \ldots, T-1$,

$$\begin{aligned}\frac{\eta}{2G}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] &\leq \phi(\theta^{(0)}) - \mathbb{E}[\phi(\theta^{(T)})] + \frac{\eta G}{2}\sum_{t=0}^{T-1} Q_t + \frac{M_\phi\eta^2 T}{2}D_{\mathrm{ema}}^2 \\ &\leq \Phi_0 + \frac{\eta G}{2}\sum_{t=0}^{T-1} Q_t + \frac{M_\phi\eta^2 T}{2}D_{\mathrm{ema}}^2.\end{aligned}$$

Multiplying by $2G/(\eta T)$ proves Eq. (54). The minimum of the T expected squared gradient norms is no greater than their average. □

A.13.7 Coupled Tracking and Stationarity

Theorem 2 bounds stationarity in terms of the bias sequence Q_t. By coupling EMA tracking errors with model movement, we obtain an $\mathcal{O}(T^{-1/2})$ guarantee when the step size and common EMA parameter are chosen for the prescribed horizon.

For the following result, use the same initialization and input conditions as Section A.13.1, with $\beta_u = \beta_v = \beta_d = \beta \in (0, 1]$ and $\eta > 0$ fixed within a run. Define the actual direction error and deterministic second moments by

$$\begin{aligned}e^{(t)} &:= \hat{d}^{(t)} - \frac{1}{G}\nabla\phi(\theta^{(t)}), \\ \mathsf{U}_t &:= \sum_g p_g\mathbb{E}[|u_{g,1}^{(t)} - L_g(\theta^{(t)})|^2], \\ \mathsf{N}_t &:= \mathbb{E}[|v^{(t)} - Z(\theta^{(t)})|^2], \\ \mathsf{D}_t &:= \mathbb{E}[\|\hat{d}^{(t)}\|^2], \\ \mathsf{E}_t &:= \mathbb{E}[\|e^{(t)}\|^2].\end{aligned}$$

For every integer $T \geq 1$, write

$$\begin{aligned}
\bar{\mathsf{U}}_T &:= \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{U}_t, \\
\bar{\mathsf{N}}_T &:= \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{N}_t, \\
\bar{\mathsf{D}}_T &:= \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{D}_t, \\
\bar{\mathsf{E}}_T &:= \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{E}_t, \\
\bar{\mathsf{G}}_T &:= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2].
\end{aligned}$$

Proposition A.5 (Coupled EMA tracking bound). In the preceding setting, for every integer $T \geq 1$,

$$\bar{\mathsf{E}}_T \leq \frac{c_0}{\beta T} + c_1\beta + c_2\left(\frac{\eta}{\beta}\right)^2\left(\frac{\bar{\mathsf{G}}_T}{G^2} + \bar{\mathsf{E}}_T\right), \tag{55}$$

where the following nonnegative constants do not depend on T, η, β:

$$\begin{aligned}
c_0 &:= \frac{2H^2}{G^2} + \frac{H^2\sigma_\ell^2}{G^2\lambda^2}\left(6e^{4B_L/\lambda} + 18e^{8B_L/\lambda}\right), \\
c_1 &:= \frac{H^2\sigma_\ell^2}{G^2\lambda^2}\left(6e^{4B_L/\lambda} + 12e^{8B_L/\lambda}\right) + \frac{6e^{4B_L/\lambda}\sigma_\ell^2\sigma^2}{G^2\lambda^2} \\
&\quad + \frac{2e^{4B_L/\lambda}}{G^2}\left(\sigma + \frac{H\sigma_\ell}{\lambda}\right)^2, \\
c_2 &:= \frac{8M_\phi^2}{G^2} + \frac{H^4}{G^2\lambda^2}\left(12e^{4B_L/\lambda} + 48e^{8B_L/\lambda}\right).
\end{aligned} \tag{56}$$

Proof. **Initial conditions and actual model movement.** Set

$$u_{g,1}^{(-1)} := u_{g,0}^{(0)}, \quad \theta^{(-1)} := \theta^{(0)}, \quad \hat{d}^{(-1)} := 0, \quad e^{(-1)} := -\frac{1}{G}\nabla\phi(\theta^{(0)}).$$

The initial error bounds and the exponential Lipschitz bound give

$$\mathsf{U}_{-1} \leq \sigma_\ell^2, \qquad \mathsf{N}_0 \leq \frac{e^{2B_L/\lambda}}{\lambda^2}\sigma_\ell^2, \qquad \mathsf{D}_{-1} = 0, \qquad \mathsf{E}_{-1} \leq \frac{H^2}{G^2}.$$

Since $\hat{d}^{(t)} = G^{-1}\nabla\phi(\theta^{(t)}) + e^{(t)}$,

$$\bar{\mathsf{D}}_T \leq \frac{2}{G^2}\bar{\mathsf{G}}_T + 2\bar{\mathsf{E}}_T, \qquad \mathbb{E}[\|\theta^{(t+1)} - \theta^{(t)}\|^2] = \eta^2\mathsf{D}_t. \tag{57}$$

We retain this actual movement rather than replace it by $\eta^2 D_{\mathrm{ema}}^2$ in the tracking recursions.

Loss EMA time average. The conditional calculation in Lemma A.13 removes the cross term with the current loss noise. Use

$$\mathbb{E}[|L_g(\theta^{(t)}) - L_g(\theta^{(t-1)})|^2] \leq H^2\eta^2\mathsf{D}_{t-1}$$

instead of a uniform drift bound. Multiplying the client recursions by p_g and summing gives, for $t \geq 0$,

$$\mathsf{U}_t \leq (1-\beta)\mathsf{U}_{t-1} + \frac{(1-\beta)^2H^2\eta^2}{\beta}\mathsf{D}_{t-1} + \beta^2\sigma_\ell^2.$$

At $t = 0$, the model-movement term is zero; the initial error is history-measurable, so the same conditional argument applies. Summing in time,

$$\begin{aligned}\beta T\bar{\mathsf{U}}_T &\le (1-\beta)\mathsf{U}_{-1} - (1-\beta)\mathsf{U}_{T-1} + \frac{(1-\beta)^2H^2\eta^2}{\beta}\sum_{t=0}^{T-1}\mathsf{D}_{t-1} + \beta^2 T\sigma_\ell^2 \\ &\le \sigma_\ell^2 + \frac{H^2\eta^2}{\beta}T\bar{\mathsf{D}}_T + \beta^2 T\sigma_\ell^2.\end{aligned}$$

Here $\mathsf{D}_{-1} = 0$ and the omitted terminal moments are nonnegative. Thus

$$\bar{\mathsf{U}}_T \le \frac{\sigma_\ell^2}{\beta T} + \beta\sigma_\ell^2 + H^2\frac{\eta^2}{\beta^2}\bar{\mathsf{D}}_T. \tag{58}$$

Normalization EMA time average. Write

$$\delta_w^{(t)} := \sum_g(\hat{w}_{g,1}^{(t)} - w_g(\theta^{(t)})), \qquad \delta_Z^{(t)} := Z(\theta^{(t+1)}) - Z(\theta^{(t)}).$$

The exact recursion is

$$v^{(t+1)} - Z(\theta^{(t+1)}) = (1-\beta)(v^{(t)} - Z(\theta^{(t)})) + \beta\delta_w^{(t)} - \delta_Z^{(t)}.$$

Convexity of the square with weights $1-\beta, \beta$, followed by Young's inequality, gives

$$\begin{aligned}|v^{(t+1)} - Z(\theta^{(t+1)})|^2 &\le (1-\beta)|v^{(t)} - Z(\theta^{(t)})|^2 + \beta|\delta_w^{(t)} - \delta_Z^{(t)}/\beta|^2 \\ &\le (1-\beta)|v^{(t)} - Z(\theta^{(t)})|^2 + 2\beta|\delta_w^{(t)}|^2 + \frac{2}{\beta}|\delta_Z^{(t)}|^2.\end{aligned}$$

The exponential Lipschitz bound, Jensen's inequality with weights p_g, and the $He^{B_L/\lambda}/\lambda$-Lipschitz property of Z yield

$$\mathbb{E}[|\delta_w^{(t)}|^2] \le \frac{e^{2B_L/\lambda}}{\lambda^2}\mathsf{U}_t, \qquad \mathbb{E}[|\delta_Z^{(t)}|^2] \le \frac{H^2e^{2B_L/\lambda}}{\lambda^2}\eta^2\mathsf{D}_t.$$

Consequently,

$$\mathsf{N}_{t+1} \le (1-\beta)\mathsf{N}_t + \frac{2\beta e^{2B_L/\lambda}}{\lambda^2}\mathsf{U}_t + \frac{2H^2e^{2B_L/\lambda}\eta^2}{\beta\lambda^2}\mathsf{D}_t.$$

The time sum has boundary contribution $\mathsf{N}_0 - \mathsf{N}_T$. Dropping $-\mathsf{N}_T$ and using the initial bound gives

$$\bar{\mathsf{N}}_T \le \frac{e^{2B_L/\lambda}\sigma_\ell^2}{\lambda^2\beta T} + \frac{2e^{2B_L/\lambda}}{\lambda^2}\bar{\mathsf{U}}_T + \frac{2H^2e^{2B_L/\lambda}}{\lambda^2}\frac{\eta^2}{\beta^2}\bar{\mathsf{D}}_T.$$

Substituting Eq. (58),

$$\bar{\mathsf{N}}_T \le \frac{3e^{2B_L/\lambda}\sigma_\ell^2}{\lambda^2\beta T} + \frac{2e^{2B_L/\lambda}\sigma_\ell^2}{\lambda^2}\beta + \frac{4H^2e^{2B_L/\lambda}}{\lambda^2}\frac{\eta^2}{\beta^2}\bar{\mathsf{D}}_T. \tag{59}$$

Current-input covariance. Retain the coefficient error $c_g^{(t)}$ from Lemma A.16. Its pointwise bound is

$$|c_g^{(t)}| \le \frac{p_g e^{2B_L/\lambda}}{\lambda}|u_{g,1}^{(t)} - L_g(\theta^{(t)})| + p_g e^{3B_L/\lambda}|v^{(t)} - Z(\theta^{(t)})|.$$

Given $\mathcal{F}^{(t)}$, consider

$$f_{g,t}(x) := \frac{p_g}{v^{(t)}}\exp\left(\frac{(1-\beta)u_{g,0}^{(t)} + \beta x}{\lambda}\right) - \xi_g(\theta^{(t)}).$$

Then $c_g^{(t)} = f_{g,t}(\bar{\ell}_g^{(t)})$ and $|f'_{g,t}(x)| \le p_g\beta e^{2B_L/\lambda}/\lambda$ on $[-B_L, B_L]$. Since $L_g(\theta^{(t)})$ is history-measurable, the least-squares property of conditional expectation gives

$$\begin{aligned}\mathrm{Var}(c_g^{(t)} \mid \mathcal{F}^{(t)}) &\le \mathbb{E}[|f_{g,t}(\bar{\ell}_g^{(t)}) - f_{g,t}(L_g(\theta^{(t)}))|^2 \mid \mathcal{F}^{(t)}] \\ &\le \frac{p_g^2\beta^2e^{4B_L/\lambda}}{\lambda^2}\sigma_\ell^2.\end{aligned}$$

Set $\varepsilon_g^{(t)} := \widehat{\nabla L_g}^{(t)} - \nabla L_g(\theta^{(t)})$. Its conditional mean is zero, so

$$\mathbb{E}[c_g^{(t)}\widehat{\nabla L_g}^{(t)} \mid \mathcal{F}^{(t)}] = \mathbb{E}[c_g^{(t)} \mid \mathcal{F}^{(t)}]\nabla L_g(\theta^{(t)}) + \mathbb{E}[(c_g^{(t)} - \mathbb{E}[c_g^{(t)} \mid \mathcal{F}^{(t)}])\varepsilon_g^{(t)} \mid \mathcal{F}^{(t)}].$$

Conditional Cauchy–Schwarz bounds the covariance without independence:

$$\begin{aligned}&\|\mathbb{E}[(c_g^{(t)} - \mathbb{E}[c_g^{(t)} \mid \mathcal{F}^{(t)}])\varepsilon_g^{(t)} \mid \mathcal{F}^{(t)}]\| \\ &\quad\le \sqrt{\mathrm{Var}(c_g^{(t)} \mid \mathcal{F}^{(t)})}\sqrt{\mathbb{E}[\|\varepsilon_g^{(t)}\|^2 \mid \mathcal{F}^{(t)}]} \le \frac{p_g\beta e^{2B_L/\lambda}}{\lambda}\sigma_\ell\sigma.\end{aligned}$$

Actual input mean and centered noise. Define the input to the aggregate direction EMA by

$$Y_t := \frac{1}{G}\sum_g \frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\widehat{\nabla L_g}^{(t)}.$$

The exact-weight noise has conditional mean zero, hence

$$\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)}) = \frac{1}{G}\sum_g \mathbb{E}[c_g^{(t)}\widehat{\nabla L_g}^{(t)} \mid \mathcal{F}^{(t)}].$$

Conditional Jensen's inequality and the preceding covariance bound imply

$$\left\|\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|_{L^2} \le \frac{H}{G}\sum_g \|c_g^{(t)}\|_{L^2} + \frac{\beta e^{2B_L/\lambda}}{G\lambda}\sigma_\ell\sigma.$$

Minkowski's inequality and weighted Cauchy–Schwarz give

$$\sum_g \|c_g^{(t)}\|_{L^2} \le \frac{e^{2B_L/\lambda}}{\lambda}\sqrt{\mathsf{U}_t} + e^{3B_L/\lambda}\sqrt{\mathsf{N}_t}.$$

Thus, squaring the three-term bound,

$$\begin{aligned}&\mathbb{E}\left\|\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2 \\ &\quad\le \frac{3H^2e^{4B_L/\lambda}}{G^2\lambda^2}\mathsf{U}_t + \frac{3H^2e^{6B_L/\lambda}}{G^2}\mathsf{N}_t + \frac{3\beta^2e^{4B_L/\lambda}\sigma_\ell^2\sigma^2}{G^2\lambda^2}.\end{aligned} \tag{60}$$

For each client's centered input, the decomposition is

$$\begin{aligned}&\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\widehat{\nabla L_g}^{(t)} - \mathbb{E}\left[\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\widehat{\nabla L_g}^{(t)} \,\middle|\, \mathcal{F}^{(t)}\right] \\ &\quad= (c_g^{(t)} - \mathbb{E}[c_g^{(t)} \mid \mathcal{F}^{(t)}])\nabla L_g(\theta^{(t)}) + \frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\varepsilon_g^{(t)} - \mathbb{E}\left[\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\varepsilon_g^{(t)} \,\middle|\, \mathcal{F}^{(t)}\right].\end{aligned}$$

Conditional L^2 triangle inequalities and the fact that centering does not increase the second moment give

$$\begin{aligned}&\left(\mathbb{E}\left[\left\|\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\widehat{\nabla L_g}^{(t)} - \mathbb{E}\left[\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\widehat{\nabla L_g}^{(t)} \,\middle|\, \mathcal{F}^{(t)}\right]\right\|^2 \,\middle|\, \mathcal{F}^{(t)}\right]\right)^{1/2} \\ &\quad\le H\sqrt{\mathrm{Var}(c_g^{(t)} \mid \mathcal{F}^{(t)})} + \left(\mathbb{E}\left[\left|\frac{\hat{w}_{g,1}^{(t)}}{v^{(t)}}\right|^2\|\varepsilon_g^{(t)}\|^2 \,\middle|\, \mathcal{F}^{(t)}\right]\right)^{1/2} \\ &\quad\le p_g e^{2B_L/\lambda}\left(\sigma + \frac{\beta H\sigma_\ell}{\lambda}\right).\end{aligned}$$

Summing over clients by another conditional L^2 triangle inequality yields

$$\begin{aligned}\mathbb{E}[Y_t - \mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] \mid \mathcal{F}^{(t)}] &= 0,\\ \mathbb{E}[\|Y_t - \mathbb{E}[Y_t \mid \mathcal{F}^{(t)}]\|^2 \mid \mathcal{F}^{(t)}] &\le \frac{e^{4B_L/\lambda}}{G^2}\left(\sigma + \frac{\beta H \sigma_\ell}{\lambda}\right)^2\\ &\le \frac{e^{4B_L/\lambda}}{G^2}\left(\sigma + \frac{H\sigma_\ell}{\lambda}\right)^2 \quad \text{a.s.}\end{aligned} \tag{61}$$

Actual direction EMA time average. The actual aggregate obeys $\hat{d}^{(t)} = (1-\beta)\hat{d}^{(t-1)} + \beta Y_t$. Subtracting the current target,

$$\begin{aligned}e^{(t)} = (1-\beta)&\left[e^{(t-1)} - \left(\frac{1}{G}\nabla\phi(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t-1)})\right)\right]\\ &+ \beta\left(\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right) + \beta\left(Y_t - \mathbb{E}[Y_t \mid \mathcal{F}^{(t)}]\right).\end{aligned}$$

Everything except the last centered term is history-measurable. The conditional cross term therefore vanishes, giving the identity

$$\begin{aligned}&\mathbb{E}[\|e^{(t)}\|^2 \mid \mathcal{F}^{(t)}]\\ &\quad= \left\|(1-\beta)\left[e^{(t-1)} - \left(\frac{1}{G}\nabla\phi(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t-1)})\right)\right] + \beta\left(\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right)\right\|^2\\ &\qquad+ \beta^2\mathbb{E}[\|Y_t - \mathbb{E}[Y_t \mid \mathcal{F}^{(t)}]\|^2 \mid \mathcal{F}^{(t)}].\end{aligned}$$

Convexity first bounds the squared norm on the right by

$$\begin{aligned}(1-\beta)&\left\|e^{(t-1)} - \left(\frac{1}{G}\nabla\phi(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t-1)})\right)\right\|^2\\ &+ \beta\left\|\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2.\end{aligned}$$

For $0 < \beta < 1$, apply parametric Young's inequality to the first squared norm with parameter $\beta/[2(1-\beta)]$. This gives

$$\begin{aligned}&\left(1 - \frac{\beta}{2}\right)\|e^{(t-1)}\|^2 + \frac{(1-\beta)(2-\beta)}{\beta}\left\|\frac{1}{G}\nabla\phi(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t-1)})\right\|^2\\ &\quad+ \beta\left\|\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2\end{aligned}$$

as an upper bound. For $\beta = 1$, only the conditional input-mean error remains in the original squared norm, and the same bound holds. Smoothness and the actual model update imply

$$\mathbb{E}\left[\left\|\frac{1}{G}\nabla\phi(\theta^{(t)}) - \frac{1}{G}\nabla\phi(\theta^{(t-1)})\right\|^2\right] \le \frac{M_\phi^2\eta^2}{G^2}\mathsf{D}_{t-1},$$

with both sides zero for $t = 0$. Taking full expectations, using $(1-\beta)(2-\beta) \le 2$, and applying Eq. (61),

$$\begin{aligned}\mathsf{E}_t \le &\left(1 - \frac{\beta}{2}\right)\mathsf{E}_{t-1} + \frac{2M_\phi^2\eta^2}{\beta G^2}\mathsf{D}_{t-1}\\ &+ \beta\mathbb{E}\left[\left\|\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2\right] + \frac{\beta^2 e^{4B_L/\lambda}}{G^2}\left(\sigma + \frac{H\sigma_\ell}{\lambda}\right)^2.\end{aligned}$$

Time summation, followed by dropping the nonpositive terminal contribution and using $\mathsf{E}_{-1} \le H^2/G^2$, yields

$$\begin{aligned}\bar{\mathsf{E}}_T \le &\frac{2H^2}{G^2\beta T} + \frac{4M_\phi^2}{G^2}\frac{\eta^2}{\beta^2}\bar{\mathsf{D}}_T\\ &+ \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathbb{E}[Y_t \mid \mathcal{F}^{(t)}] - \frac{1}{G}\nabla\phi(\theta^{(t)})\right\|^2\right] + \frac{2\beta e^{4B_L/\lambda}}{G^2}\left(\sigma + \frac{H\sigma_\ell}{\lambda}\right)^2.\end{aligned} \tag{62}$$

Substitution and coupling. Substitute Eq. (60) and then Eqs. (58)–(59) into Eq. (62). Using $\beta^2 \le \beta$ for the covariance-squared term and collecting the explicit constants in Eq. (56) gives

$$\bar{\mathsf{E}}_T \le \frac{c_0}{\beta T} + c_1\beta + \frac{c_2}{2}\frac{\eta^2}{\beta^2}\bar{\mathsf{D}}_T.$$

Finally, apply Eq. (57) to obtain Eq. (55). □

Corollary A.1 (Stationarity with horizon-dependent EMA parameters). Under Proposition A.5, suppose

$$\kappa := c_2\left(\frac{\eta}{\beta}\right)^2 < \frac{1}{2}.$$

With Φ_0 and D_{ema}^2 as in Theorem 2, for every integer $T \ge 1$,

$$\begin{aligned}\min_{0\le t<T}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] \le \bar{\mathsf{G}}_T \le \frac{1-\kappa}{1-2\kappa}\left(\frac{2G\Phi_0}{\eta T} + M_\phi G\eta D_{\mathrm{ema}}^2\right)\\ + \frac{G^2}{1-2\kappa}\left(\frac{c_0}{\beta T} + c_1\beta\right).\end{aligned} \tag{63}$$

In particular, choose constants $\eta_0 > 0$ and $0 < \beta_0 \le 1$ independent of T, with $c_2(\eta_0/\beta_0)^2 < 1/2$. Before a run of prescribed horizon T, set

$$\eta = \frac{\eta_0}{\sqrt{T}}, \qquad \beta_u = \beta_v = \beta_d = \beta = \frac{\beta_0}{\sqrt{T}},$$

and hold them fixed throughout that run. Then $\min_{0\le t<T}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] \le \bar{\mathsf{G}}_T = \mathcal{O}(T^{-1/2})$.

Proof. Since $b^{(t)} = \mathbb{E}[e^{(t)} \mid \mathcal{F}^{(t)}]$, conditional Jensen's inequality and the tower property give $\mathbb{E}\|b^{(t)}\|^2 \le \mathsf{E}_t$. Take full expectations in Lemma A.17, using this bound and Lemma A.12, to obtain

$$\frac{\eta}{2G}\mathbb{E}[\|\nabla\phi(\theta^{(t)})\|^2] \le \mathbb{E}[\phi(\theta^{(t)})] - \mathbb{E}[\phi(\theta^{(t+1)})] + \frac{\eta G}{2}\mathsf{E}_t + \frac{M_\phi\eta^2}{2}D_{\mathrm{ema}}^2.$$

Summing in time and multiplying by $2G/(\eta T)$,

$$\bar{\mathsf{G}}_T \le \frac{2G\Phi_0}{\eta T} + G^2\bar{\mathsf{E}}_T + M_\phi G\eta D_{\mathrm{ema}}^2. \tag{64}$$

Proposition A.5 and $\kappa < 1/2$ imply

$$\bar{\mathsf{E}}_T \le \frac{1}{1-\kappa}\left(\frac{c_0}{\beta T} + c_1\beta\right) + \frac{\kappa}{(1-\kappa)G^2}\bar{\mathsf{G}}_T.$$

Substituting into Eq. (64) and collecting the gradient terms gives

$$\frac{1-2\kappa}{1-\kappa}\bar{\mathsf{G}}_T \le \frac{2G\Phi_0}{\eta T} + \frac{G^2}{1-\kappa}\left(\frac{c_0}{\beta T} + c_1\beta\right) + M_\phi G\eta D_{\mathrm{ema}}^2.$$

Multiplication by $(1-\kappa)/(1-2\kappa)$ proves Eq. (63); the minimum is at most the average.

For the prescribed horizon-dependent choices, $\kappa = c_2(\eta_0/\beta_0)^2 < 1/2$ is independent of T, and $0 < \beta \le 1$ for every $T \ge 1$. Substitution gives the explicit bound

$$\begin{aligned}\bar{\mathsf{G}}_T &\le \frac{1}{\sqrt{T}}\left[\frac{1-\kappa}{1-2\kappa}\left(\frac{2G\Phi_0}{\eta_0} + M_\phi G\eta_0 D_{\mathrm{ema}}^2\right) + \frac{G^2}{1-2\kappa}\left(\frac{c_0}{\beta_0} + c_1\beta_0\right)\right]\\ &= \mathcal{O}(T^{-1/2}).\end{aligned} \tag{65}$$

□

Algorithm 6 One Federated Round of DRO-KL (EMA-Extended)

Require: Current global model $\theta^{(t)}$, current normalization constant $v^{(t)}$, set of participating clients $S^{(t)} \subset \{1, \dots, G\}$, temperature λ, learning rate η, number of local steps K, EMA parameters $\beta_u, \beta_v, \beta_d$

Ensure: Updated global model $\theta^{(t+1)}$, updated normalization constant $v^{(t+1)}$

1: Server broadcasts $(\theta^{(t)}, v^{(t)})$ to all clients in $S^{(t)}$
2: **for all** clients $g \in S^{(t)}$ **in parallel do**
3: $(\theta_g^{(t+1)}, \tilde{v}_g^{(t)}) \leftarrow \text{LocalUpdateEMA}(g, \theta^{(t)}, v^{(t)})$ ▷ Algorithm 1
4: **end for**
5: Server collects $\{(\theta_g^{(t+1)}, \tilde{v}_g^{(t)})\}_{g \in S^{(t)}}$
6: $(\theta^{(t+1)}, v^{(t+1)}) \leftarrow \text{ServerAggregateEMA}(\{\theta_g^{(t+1)}\}, \{\tilde{v}_g^{(t)}\})$ ▷ Algorithm 8
7: **return** $\theta^{(t+1)}$, $v^{(t+1)}$

Scope and interpretation. Corollary A.1 applies to a family of runs whose parameters are chosen before each prescribed horizon, not to arbitrary fixed EMA coefficients or to a decreasing schedule within a single run. In particular, it does not give the same rate for the fixed experimental settings. Similar horizon-dependent choices of the step size and EMA parameters are used in the convergence analysis of Guo and Yang [2024].

The guarantee concerns the time average or minimum of expected squared gradient norms. It does not imply convergence of the model sequence, global optimality, or monotone improvement at every round.

Algorithms 6–8 give the expanded pseudocode for Algorithm 1. The analysis above is restricted to full participation and $K = 1$.

A.14 Discussion of Convergence Parameters

Effect of λ and K. Smaller λ yields greater robustness but increases M_ϕ (Proposition 2), requiring smaller learning rates; $\lambda \in \{2, 5\}$ provides the best trade-off in our experiments. Multiple local steps ($K > 1$) introduce client drift $\mathcal{O}(\eta^2 K^2 M_\phi^2)$; we use $K{=}1$ in all main experiments (see Appendix C.2 for an ablation over $K \in \{1, 2, 3\}$). The general-K drift result concerns the idealized exact-weight method; the EMA guarantee is restricted to $K = 1$ and full participation. Our theory assumes vanilla SGD. Momentum and weight decay require additional analysis and are not covered by these guarantees.

Client-sampling variability. For $G > 1$, the conditional variance of the sampled gradient mean is bounded by $(G - m)\zeta^2/(m(G - 1))$ under Assumptions 4 and 6. This sampling deviation has zero conditional mean and is not an irreducible optimization floor. Theorem 1 therefore has a vanishing bound for general K, and Corollary B.1 removes local drift entirely.

B Additional Theoretical Results

B.1 GroupDRO Weight Collapse under Exponentiated Gradient

The following proposition formalizes the weight-collapse phenomenon observed empirically (Section 6, Figure 4b): the exponentiated-gradient update used by GroupDRO [Sagawa et al., 2020] concentrates mixture weights on a single group, whereas KL-regularized weights admit a uniform lower bound. The analysis is conducted at the population level (using the true expected losses $L_g(\theta)$ rather than minibatch or EMA estimates) and assumes a persistent worst-case group with a strict loss gap. The proposition therefore serves as a *qualitative* explanation of the collapse mechanism; extending its rate to normalized or EMA-based estimates would require additional assumptions and analysis.

Relation to the GroupDRO implementation. The federated training code updates retained client loss EMAs and, in its divisive branch, subtracts one after dividing each estimate by their common

Algorithm 7 Local Update for Client g (DRO-KL, EMA-Extended) — expanded pseudocode

Require: Global model parameters $\theta^{(t)}$, normalization constant $v^{(t)}$, local dataset $\mathcal{D}_g$, reference weight p_g, temperature λ, learning rate η, number of local steps K, EMA parameters $\beta_u, \beta_v, \beta_d \in (0, 1]$, batch size B, EMA states carried over from the previous round: $u_{g,0}^{(t)}$, $\tilde{v}_{g,0}^{(t)}$, $\hat{d}_{g,0}^{(t)}$ (at t=0 initialize with $u_{g,0}^{(0)} \approx L_g(\theta^{(0)})$ via a minibatch estimate (the analysis uses the initialization conditions of Section A.13.1), $\tilde{v}_{g,0}^{(0)} = p_g e^{u_{g,0}^{(0)}/\lambda}$, $\hat{d}_{g,0}^{(0)} = \mathbf{0}$; for $t \geq 1$ set $u_{g,0}^{(t)} = u_{g,K}^{(t-1)}$, $\tilde{v}_{g,0}^{(t)} = \tilde{v}_{g,K}^{(t-1)}$, $\hat{d}_{g,0}^{(t)} = \hat{d}_{g,K}^{(t-1)}$)

Ensure: Updated local model $\theta_g^{(t+1)}$, EMA normalization estimate $\tilde{v}_{g,K}^{(t)}$

$\theta_{g,0}^{(t)} \leftarrow \theta^{(t)}$

for $k = 1, \ldots, K$ **do**

  Sample a minibatch $x_i \sim \mathcal{D}_g,\ i = 1, \ldots, B$

  Sample trajectories $\tau_i \sim \pi_{\theta_{g,k-1}^{(t)}}(\cdot \mid x_i),\ i = 1, \ldots, B$

  Compute losses $\ell_i = \ell(\tau_i),\ i = 1, \ldots, B$

  Compute baseline b

  Estimate policy gradient:

$$\widehat{\nabla L_g} \leftarrow \frac{1}{B} \sum_{i=1}^{B} (\ell_i - b) \nabla_\theta \log \pi_{\theta_{g,k-1}^{(t)}}(\tau_i \mid x_i)$$

  Compute batch mean loss: $\bar{\ell}_g \leftarrow \frac{1}{B} \sum_{i=1}^{B} \ell_i$

  Update loss EMA: $u_{g,k}^{(t)} \leftarrow (1 - \beta_u)\, u_{g,k-1}^{(t)} + \beta_u\, \bar{\ell}_g$

  Compute KL-DRO weight: $\hat{w}_{g,k}^{(t)} \leftarrow p_g \exp(u_{g,k}^{(t)}/\lambda)$

  Update normalization EMA: $\tilde{v}_{g,k}^{(t)} \leftarrow (1 - \beta_v)\, \tilde{v}_{g,k-1}^{(t)} + \beta_v\, \hat{w}_{g,k}^{(t)}$

  Update gradient EMA: $\hat{d}_{g,k}^{(t)} \leftarrow (1 - \beta_d)\, \hat{d}_{g,k-1}^{(t)} + \beta_d\, (\hat{w}_{g,k}^{(t)}/v^{(t)})\, \widehat{\nabla L_g}$

  Local model update: $\theta_{g,k}^{(t)} \leftarrow \theta_{g,k-1}^{(t)} - \eta\, \hat{d}_{g,k}^{(t)}$

end for

$\theta_g^{(t+1)} \leftarrow \theta_{g,K}^{(t)}$

return $\theta_g^{(t+1)},\ \tilde{v}_{g,K}^{(t)}$

mean. It passes these values to a multiplicative-weight routine once per round, rather than using the true expected losses in the proposition. For the main-result GroupDRO runs, the recorded setting is $\beta_u = 0.05$ with full participation. The current q weights client gradients during a round; after client processing, the implementation updates q once using the vector of EMA-normalized losses, and that updated vector is used in the next round. Under partial participation, unselected clients retain their previous EMA values and those values remain in the round-end update. When the mean is positive and the numerical guards are inactive, the common shift cancels in normalization, while division by the mean changes the effective step size to $\eta_q/\bar{u}$. This scale changes as the retained estimates change. The weight-update routine also clips exponent arguments and imposes a lower bound before normalization; the training source uses zero normalized inputs when the mean is too small. These operations, as well as loss-estimation noise and EMA history, are outside the raw-loss recurrence analyzed below.

The following restates Proposition 1 with full details before giving the proof. Consider G groups with losses $L_1(\theta), \ldots, L_G(\theta)$, a reference distribution $p \in \Delta_G$ with $p_g > 0$ for all g, and the exponentiated-gradient simplex update used by GroupDRO [Sagawa et al., 2020]—*not* the KL-regularized update of Algorithms 2–1—:

$$q_g^{(t+1)} = \frac{q_g^{(t)} \exp\big(\eta_q\, L_g(\theta^{(t)})\big)}{\sum_{g'=1}^{G} q_{g'}^{(t)} \exp\big(\eta_q\, L_{g'}(\theta^{(t)})\big)}, \qquad q_g^{(0)} = p_g.$$

Let $g^\star = \arg\max_g L_g(\theta^{(t)})$ be the highest-loss group at each round and define the minimum loss gap $\delta_{\min} := \min_{0 \leq t \leq T-1} \big(L_{g^\star}(\theta^{(t)}) - \max_{g \neq g^\star} L_g(\theta^{(t)})\big) > 0$. If $g^\star$ is the same group for all

Algorithm 8 Server Aggregation for DRO-KL (EMA-Extended)

Require: Client models $\{\theta_g^{(t+1)}\}_{g\in S^{(t)}}$, client EMA normalization estimates $\{\tilde{v}_{g,K}^{(t)}\}_{g\in S^{(t)}}$
Ensure: Updated global model $\theta^{(t+1)}$, updated normalization constant $v^{(t+1)}$
1: Aggregate models (FedAvg):

$$\theta^{(t+1)} \leftarrow \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}} \theta_g^{(t+1)}$$

2: Aggregate EMA normalization estimates:

$$v^{(t+1)} \leftarrow \frac{G}{|S^{(t)}|}\sum_{g\in S^{(t)}} \tilde{v}_{g,K}^{(t)}$$

3: Broadcast $\theta^{(t+1)}$ and $v^{(t+1)}$ to clients
4: **return** $\theta^{(t+1)}$, $v^{(t+1)}$

rounds, then after T rounds:

$$q_{g^\star}^{(T)} \;\geq\; 1 - \frac{1-p_{g^\star}}{p_{g^\star}}\exp\bigl(-\eta_q\,\delta_{\min}\,T\bigr).$$

Proof. Define the log-ratio $r_g^{(t)} := \log q_{g^\star}^{(t)} - \log q_g^{(t)}$ for any $g \neq g^\star$. By the update rule,

$$r_g^{(t+1)} = r_g^{(t)} + \eta_q\bigl(L_{g^\star}(\theta^{(t)}) - L_g(\theta^{(t)})\bigr) \geq r_g^{(t)} + \eta_q\delta_{\min},$$

since $L_{g^\star}(\theta^{(t)}) - L_g(\theta^{(t)}) \geq \delta_{\min}$ by assumption. Telescoping from $r_g^{(0)} = \log(p_{g^\star}/p_g)$ (since $q_g^{(0)} = p_g$):

$$r_g^{(T)} \geq \log\frac{p_{g^\star}}{p_g} + \eta_q\,\delta_{\min}\,T \quad\Longrightarrow\quad \frac{q_{g^\star}^{(T)}}{q_g^{(T)}} \geq \frac{p_{g^\star}}{p_g}\exp(\eta_q\,\delta_{\min}\,T).$$

Since $\sum_{g=1}^G q_g^{(T)} = 1$,

$$q_{g^\star}^{(T)} = \frac{1}{1+\sum_{g\neq g^\star} q_g^{(T)}/q_{g^\star}^{(T)}} \geq \frac{1}{1+\frac{1-p_{g^\star}}{p_{g^\star}}\exp(-\eta_q\,\delta_{\min}\,T)} \geq 1 - \frac{1-p_{g^\star}}{p_{g^\star}}\exp(-\eta_q\,\delta_{\min}\,T),$$

where the second inequality uses $\sum_{g\neq g^\star}(p_g/p_{g^\star}) = (1-p_{g^\star})/p_{g^\star}$, and the last step uses $1/(1+x) \geq 1-x$ for $x \geq 0$. □

In particular, setting the right-hand side to ϵ and solving for T shows that $q_{g^\star}^{(T)} \geq 1-\epsilon$ whenever $T \geq \frac{1}{\eta_q\delta_{\min}}\log\Bigl(\frac{1-p_{g^\star}}{p_{g^\star}\,\epsilon}\Bigr)$; that is, the weight concentrates on a single group at an exponential rate in T.

Remark B.1 (On the assumptions). The two key assumptions—that $g^\star$ is the same group for all rounds and that $\delta_{\min} > 0$—are sufficient conditions for exponentially fast collapse. The strict positivity $\delta_{\min} > 0$ ensures that $g^\star$ is the *unique* maximizer at every round; if two groups tied for the highest loss at some round, the gap at that round would be zero. Absence of exact ties alone does not ensure a fixed maximizer across rounds or a gap bounded away from zero as the horizon grows. The proposition does not assert inevitable concentration when the highest-loss group changes or the gap vanishes.

Remark B.2 (KL-regularized weights avoid cumulative collapse). In contrast, the KL-regularized objective $\phi(\theta) = \lambda\log\bigl(\sum_g p_g\exp(L_g(\theta)/\lambda)\bigr)$ yields closed-form softmax weights $\xi_g(\theta) = p_g\exp(L_g(\theta)/\lambda)\big/\sum_{g'} p_{g'}\exp(L_{g'}(\theta)/\lambda)$. If all losses are bounded in $[0, B_L]$ (Assumption 1), then for every group g:

$$\xi_g(\theta) \;\geq\; p_{\min}\exp\left(-\frac{B_L}{\lambda}\right), \qquad p_{\min} := \min_g p_g.$$

For positive $p_{\min}$ and finite $\lambda > 0$, this lower bound is strictly positive in exact arithmetic and uniform across rounds. It can nevertheless be numerically negligible when B_L/λ is large, and does not ensure substantial weight for every group. The exact KL weights depend on the current losses and fixed reference p, not on multiplicative accumulation of previous weights. The temperature λ controls the interpolation between reference-distribution weighting (large λ; $\xi_g \approx p_g$) and worst-case concentration (small λ; ξ_g peaks at the highest-loss group). The C11 additive-centering experiment at $\lambda = 2$ (Appendix A.3) illustrates the latter behavior.

B.2 Tighter Convergence Rate for $K = 1$

Theorem 1 has no constant heterogeneity floor under Assumption 6, including for general K. For $K = 1$, client drift vanishes and the aggregate is conditionally unbiased for $\nabla\phi(\theta^{(t)})/G$ after averaging over client selection. A direct variance calculation gives a tighter second-moment bound without Assumption 4.

Corollary B.1 (Convergence for $K = 1$ without heterogeneity floor). Under Assumptions 2, 3, 5, and 6 (i.e., excluding Assumption 4), with $K = 1$ and $\Phi_0 > 0$, set the step size $\eta = \sqrt{2\Phi_0/(M_\phi T\tilde{\sigma}^2)}$ where $\tilde{\sigma}^2 := \sigma^2/m + H^2$, and assume $T \geq 2M_\phi\Phi_0/(G^2\tilde{\sigma}^2)$, so that $\eta \leq G/M_\phi$. Then Algorithm 2 satisfies

$$\min_{0\leq t<T} \mathbb{E}\big[\|\nabla\phi(\theta^{(t)})\|^2\big] \;\leq\; 2G\sqrt{\frac{2M_\phi\,\Phi_0\,\tilde{\sigma}^2}{T}}.$$

The expectation includes client selection and local minibatch randomness. All terms vanish as $T \to \infty$, at rate $\mathcal{O}(T^{-1/2})$.

Proof. With $K = 1$, each client $g \in S^{(t)}$ computes a single stochastic gradient $\hat{d}_g^{(t)}$ at the common global model $\theta^{(t)}$ and the server aggregates:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{|S^{(t)}|}\sum_{g\in S^{(t)}} \hat{d}_g^{(t)}.$$

There is no client drift (all clients start from $\theta^{(t)}$ and take one step), so we analyze the aggregated update directly.

Step 1: Smoothness descent. By the descent lemma (Lemma A.6):

$$\phi(\theta^{(t+1)}) \leq \phi(\theta^{(t)}) - \eta\,\big\langle\nabla\phi(\theta^{(t)}),\, \bar{d}^{(t)}\big\rangle + \frac{M_\phi\eta^2}{2}\|\bar{d}^{(t)}\|^2, \tag{66}$$

where $\bar{d}^{(t)} := \frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}} \hat{d}_g^{(t)}$.

Step 2: Bias and variance of $\bar{d}^{(t)}$. Here $\bar{d}^{(t)}$ is the same direction as $\hat{d}^{(t)}$ in Eq. (29) with $K = 1$. The client-drift term $b_{\text{drift}}^{(t)}$ is zero, but $b_{\text{het}}^{(t)}$ need not vanish for a fixed sampled subset. Its expectation conditional on $\mathcal{F}^{(t)}$ is zero.

By the definition of the local DRO gradient estimator in Eq. (6), $\hat{d}_g^{(t)} = \xi_g(\theta^{(t)})\,\widehat{\nabla L_g}$, each $\hat{d}_g^{(t)}$ is an unbiased estimate of $\nabla\phi_g(\theta^{(t)}) = \xi_g(\theta^{(t)})\nabla L_g(\theta^{(t)})$ conditioned on $\mathcal{F}^{(t)}$ and $S^{(t)}$ (by Proposition A.1). The true gradient decomposes as $\nabla\phi(\theta^{(t)}) = \sum_{g=1}^{G}\nabla\phi_g(\theta^{(t)})$.

For the participating subset $S^{(t)}$,

$$\mathbb{E}\left[\frac{1}{m}\sum_{g\in S^{(t)}}\nabla\phi_g(\theta^{(t)})\;\middle|\;\mathcal{F}^{(t)}\right] = \frac{1}{G}\sum_{g=1}^{G}\nabla\phi_g(\theta^{(t)}) = \frac{1}{G}\nabla\phi(\theta^{(t)}),$$

by Assumption 6. Hence $\mathbb{E}[\bar{d}^{(t)} \mid \mathcal{F}^{(t)}] = \nabla\phi(\theta^{(t)})/G$.

For the second moment, the conditional law of total variance gives

$$\mathrm{Var}(\bar{d}^{(t)} \mid \mathcal{F}^{(t)}) = \underbrace{\mathbb{E}\Big[\mathrm{Var}(\bar{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}) \;\Big|\; \mathcal{F}^{(t)}\Big]}_{\text{(A): within-client stochastic noise}} + \underbrace{\mathrm{Var}\Big(\mathbb{E}[\bar{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] \;\Big|\; \mathcal{F}^{(t)}\Big)}_{\text{(B): client-sampling variance}}. \tag{67}$$

Term (A). Conditioning on $\mathcal{F}^{(t)}$ and $S^{(t)}$, the local stochastic gradients are independent across clients by Assumption 6. The history-conditional variance bound in the proof of Proposition A.2 gives

$$\mathrm{Var}(\bar{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}) = \frac{1}{m^2} \sum_{g \in S^{(t)}} \mathrm{Var}(\hat{d}_g^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}) \le \frac{\sigma^2}{m}.$$

This bound does not depend on $S^{(t)}$, hence $(\mathrm{A}) \le \sigma^2/m$.

Term (B). Taking the conditional expectation over the local noise gives $\mathbb{E}[\bar{d}^{(t)} \mid \mathcal{F}^{(t)}, S^{(t)}] = m^{-1} \sum_{g \in S^{(t)}} \nabla\phi_g(\theta^{(t)})$. This is a sample mean from the finite population $\{\nabla\phi_1, \dots, \nabla\phi_G\}$. For $G > 1$, the finite-population identity proved in Lemma A.9 gives

$$\begin{aligned}(\mathrm{B}) &= \frac{G-m}{m(G-1)} \frac{1}{G} \sum_{g=1}^{G} \left\| \nabla\phi_g(\theta^{(t)}) - \frac{1}{G} \nabla\phi(\theta^{(t)}) \right\|^2 \\ &\le \frac{G-m}{m(G-1)} \frac{1}{G} \sum_{g=1}^{G} \|\nabla\phi_g(\theta^{(t)})\|^2 \le H^2.\end{aligned}$$

Here subtracting the population mean decreases the mean square, $\|\nabla\phi_g\| \le H$, and $(G-m)/(m(G-1)) \le 1$. For $G = m = 1$, $(\mathrm{B}) = 0$. This uses the finite-population identity, not Assumption 4; if that assumption is imposed, the alternative bound $(\mathrm{B}) \le (G-m)\zeta^2/(m(G-1))$ is also available.

Combining (A) and (B) with the bias–variance decomposition $\mathbb{E}[\|\bar{d}^{(t)}\|^2 \mid \mathcal{F}^{(t)}] = \|\mathbb{E}[\bar{d}^{(t)} \mid \mathcal{F}^{(t)}]\|^2 + \mathrm{Var}(\bar{d}^{(t)} \mid \mathcal{F}^{(t)})$:

$$\begin{aligned}\mathbb{E}\Big[\|\bar{d}^{(t)}\|^2 \;\Big|\; \mathcal{F}^{(t)}\Big] &\le \frac{1}{G^2} \|\nabla\phi(\theta^{(t)})\|^2 + \frac{\sigma^2}{m} + H^2 \\ &= \frac{1}{G^2} \|\nabla\phi(\theta^{(t)})\|^2 + \tilde{\sigma}^2.\end{aligned} \tag{68}$$

Step 3: Combining. Substituting Eq. (68) into Eq. (66) and taking expectations, the inner-product term becomes $-\eta\, \mathbb{E}\langle \nabla\phi(\theta^{(t)}), \bar{d}^{(t)} \rangle = -\frac{\eta}{G}\, \mathbb{E}\|\nabla\phi(\theta^{(t)})\|^2$ (since $\mathbb{E}[\bar{d}^{(t)} \mid \mathcal{F}^{(t)}] = \nabla\phi(\theta^{(t)})/G$), and the squared-norm term is bounded by Eq. (68). Unlike the general-K proof, no Young's inequality is needed: $b_{\mathrm{drift}}^{(t)} = 0$ and $\mathbb{E}[b_{\mathrm{het}}^{(t)} \mid \mathcal{F}^{(t)}] = 0$, so the conditional mean after client selection is exactly $\nabla\phi(\theta^{(t)})/G$. This gives:

$$\mathbb{E}[\phi(\theta^{(t+1)})] \le \mathbb{E}[\phi(\theta^{(t)})] - \frac{\eta}{G}\, \mathbb{E}\Big[\|\nabla\phi(\theta^{(t)})\|^2\Big] + \frac{M_\phi \eta^2}{2} \Big(\tfrac{1}{G^2}\, \mathbb{E}\Big[\|\nabla\phi(\theta^{(t)})\|^2\Big] + \tilde{\sigma}^2 \Big).$$

For $\eta \le G/M_\phi$, the coefficient of $\mathbb{E}\big[\|\nabla\phi(\theta^{(t)})\|^2\big]$ is $-\eta/G + M_\phi\eta^2/(2G^2) = -(\eta/G)(1 - M_\phi\eta/(2G)) \le -\eta/(2G)$. Rearranging and telescoping over $t = 0, \dots, T-1$:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|\nabla\phi(\theta^{(t)})\|^2\Big] \le \frac{2G\Phi_0}{\eta T} + G M_\phi \eta\, \tilde{\sigma}^2.$$

Setting $\eta = \sqrt{2\Phi_0/(M_\phi T \tilde{\sigma}^2)}$ yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla\phi(\theta^{(t)})\|^2 \le 2G \sqrt{\frac{2M_\phi\, \Phi_0\, \tilde{\sigma}^2}{T}} = \mathcal{O}\left(G \sqrt{\frac{M_\phi\, \Phi_0\, \tilde{\sigma}^2}{T}} \right).$$

The step-size condition $\eta \le G/M_\phi$ is satisfied whenever $T \ge 2M_\phi\Phi_0/(G^2\tilde{\sigma}^2)$, as required in the statement. Note that ζ^2 (Assumption 4) is not needed for the bound using $\tilde{\sigma}^2$; see Remark B.3 below for a detailed explanation. □

Remark B.3 (Client selection and local drift). Under Assumption 6, the sampling deviation $b_{\text{het}}^{(t)}$ has zero conditional mean for every K, although it may be nonzero for a realized subset. It must be averaged over $S^{(t)}$ before applying Young's inequality. For general K, Theorem 1 applies Young's inequality only to $\bar{b}^{(t)} = \mathbb{E}[b_{\text{drift}}^{(t)} \mid \mathcal{F}^{(t)}]$, whose squared norm is $\mathcal{O}(\eta^2(K-1)^2)$. Thus there is no constant heterogeneity floor for general K either. With $K = 1$, this drift also vanishes, yielding the direct minibatch-SGD argument above [Woodworth et al., 2020].

B.3 Partial Client Participation

Our algorithms (Algorithms 2 and 1) naturally support *partial participation*, where only a random subset $S^{(t)} \subset \mathcal{G}$ of clients participates in each communication round. Partial participation is essential for scalability in federated learning [Kairouz et al., 2021], and its convergence analysis under heterogeneous data has been studied for local SGD [Stich, 2019, Li et al., 2020b]. The protocols average participating client models with $\frac{1}{|S^{(t)}|}\sum_{g\in S^{(t)}}$. The idealized analysis covers fixed-size uniform partial participation; the EMA analysis in Section A.13 specializes to $|S^{(t)}| = G$ and $K = 1$. We now discuss how this relates to the compositional optimization framework of Guo and Yang [2024], which achieves tighter convergence rates under a full-participation design.

Remark B.4 (Design trade-offs for partial participation). The FGDRO-KL algorithm of Guo and Yang [2024] is designed for settings where all N machines communicate every round—a natural assumption in data-center federated learning. In their design, each machine i maintains a local estimator v_i of the global normalization constant $g(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\exp(\ell_i(\mathbf{w})/\lambda)$, and uses it in the *denominator* of the gradient estimator:

$$\mathbf{m}_{i,t} = (1-\beta_3)\,\mathbf{m}_{i,t-1} + \beta_3\,\frac{\exp(u_{i,t}/\lambda)}{v_{i,t}}\,\nabla\ell(\mathbf{w}_{i,t-1};\mathbf{z}_{i,t}).$$

In communication rounds, v is averaged across all N machines: $\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_i$. This distributed tracking of v_i contributes to their tighter compositional convergence rate. However, adapting this design to partial participation ($|S^{(t)}| < N$) would require non-trivial modifications: if machine $i \notin S^{(t)}$, its local estimators u_i, v_i, $\mathbf{m}_i$ are not updated; when machine i next participates, v_i tracks $g(\mathbf{w}^{\text{old}})$ rather than $g(\mathbf{w}^{\text{current}})$. Since v_i appears in the *denominator* of the gradient estimator, this staleness would introduce a *multiplicative* bias (not merely additive noise), invalidating the tracking-error bound in their Lemma B.2 ($\mathbb{E}\|\bar{v}_t^r - g(\mathbf{w}_t^r)\|^2 \to 0$) on which the convergence proof relies.

Possible extensions include (a) bounding the staleness of non-participating machines' estimators (analogous to asynchronous FL analyses), or (b) periodically resetting v_i upon rejoining, which introduces additional transient bias. We note that this is not a limitation of their method—which targets (and is well-suited for) the full-participation regime—but rather a trade-off inherent in the distributed-v_i design.

Our design makes a different trade-off. The idealized Algorithm 2 assumes exact DRO weights. Under Assumption 6, its client-selection deviation has zero conditional mean. The resulting variability is controlled by $C_{\text{fed}} = \sigma^2/m + 2H^2$ in Theorem 1; for $K = 1$, Corollary B.1 gives the sharper variance calculation. These results cover every fixed $m \in \{1,\dots,G\}$ under uniform sampling without replacement, not arbitrary participation processes. Time-varying or history-dependent selection requires a separately specified sampling model and corresponding bounds.

Algorithm 1 instead broadcasts the server normalization $v^{(t)}$ while retaining client EMA states across rounds. The current broadcast does not by itself control errors in those retained states or in the normalization estimate. The idealized sampling argument therefore does not establish an EMA tracking guarantee under partial participation; those errors require their own analysis. Theorem 2 and Corollary A.1 instead control tracking under full participation and $K = 1$.

B.4 Inter-Group Gradient Cosine Similarity

We provide quantitative evidence for the near-zero inter-group gradient cosine similarity claimed in Section 6.2. At the end of each epoch we compute three cosine-similarity statistics from the *unclipped* REINFORCE gradient vectors $\nabla L_g \in \mathbb{R}^D$ ($D \approx 119$K for d=32):

1. **Minority–majority (pairwise):** $\bar{c}_{\text{min-maj}} = \frac{1}{G-1}\sum_{j=2}^{G}\cos(\nabla L_1, \nabla L_j)$, where client 1 holds the minority (grid) distribution and clients $2, \ldots, G$ hold the majority (cluster) distribution.
2. **Majority–majority (pairwise):** $\bar{c}_{\text{maj-maj}} = \frac{1}{\binom{G-1}{2}}\sum_{2\le i<j\le G}\cos(\nabla L_i, \nabla L_j)$ (estimated from 20 randomly sampled pairs per epoch for efficiency).
3. **Minority vs. majority centroid:** $c_{\text{centroid}} = \cos\big(\nabla L_1,\ \frac{1}{G-1}\sum_{j=2}^{G}\nabla L_j\big)$.

Table 4 reports the epoch-averaged statistics for condition C11 (cluster/grid, n=100, G=16, minority $1/16$, d=32, 40 epochs, seed 42).

Table 4: Inter-group gradient cosine similarity on condition C11 (40 epochs, seed 42). "Mean" and "Std" are the mean and sample SD across the 40 epochs. "Per-pair std" is the average across epochs of the within-epoch standard deviation of the 15 pairwise minority–majority cosines.

	FedAvg-ERM		FedDRO-KL (λ=2)	
Statistic	Mean	Std	Mean	Std
$\bar{c}_{\text{min-maj}}$ (pairwise)	-0.007	0.033	0.003	0.030
Per-pair std of $\cos(\nabla L_1, \nabla L_j)$	0.126	—	0.127	—
$\bar{c}_{\text{maj-maj}}$ (pairwise)	-0.003	0.035	-0.001	0.046
c_{centroid} (minority vs. avg)	-0.027	0.133	0.005	0.130

All epoch-averaged cosine means in the table are within 0.03 of zero; the largest magnitude is the ERM centroid mean (-0.027). This near-orthogonality is qualitatively different from the conflicting (negative cosine) gradients studied in multi-task learning [Yu et al., 2020, Liu et al., 2021], where gradient surgery techniques are designed to resolve inter-task conflicts. Under single-rollout REINFORCE, the high estimator variance masks any directional structure, rendering gradients uninformative about group identity. With $100\times$ variance reduction (POMO-100, Appendix C.6), a richer structure emerges: gradients are aligned early ($\cos \approx +0.5$) and become near-orthogonal after convergence ($\cos \approx -0.07$); however, DRO re-weighting is ineffective in both phases (see Table 15). The within-epoch standard deviation of individual pairwise cosines ($\approx$0.13) is substantially larger than the concentration-of-measure prediction for i.i.d. random unit vectors ($\sim 1/\sqrt{D} \approx 0.003$), reflecting shared model structure. That is, individual minority–majority pairs can exhibit non-negligible positive or negative cosine values, but these fluctuations cancel across the 15 pairs, so that the *average* still centers near zero. The similarity between ERM and FedDRO-KL columns indicates that DRO re-weighting does not alter the gradient *directions*, consistent with the no-improvement finding in Section 6.2.

Figure 1 visualises these statistics as epoch-progression curves for three methods (FedAvg-ERM, FedDRO-KL λ=2, GroupDRO η_q=0.01) across three seeds on condition C7 (cluster/grid, n=50, 20 epochs). Panel (a) shows the pairwise mean $\bar{c}_{\text{min-maj}}$ with ± 1 seed standard deviation; panel (b) shows the same mean with ± 1 per-pair standard deviation. The three-seed mean remains within ± 0.05 at every epoch. Linear trends in these mean curves are not statistically significant ($p = 0.09$–0.13). Across methods, the seed-level sample SD averaged over epochs (shaded bands in panel a) is 1.3–2.6 times the sample SD across epochs of the three-seed mean curve, so seed variation is comparable to or larger than epoch-to-epoch variation.

Figure 2 extends the analysis to 40 epochs (seed 42, ERM and FedDRO-KL only) to verify that the orthogonality persists over a longer training horizon. Panel (a) plots the pairwise mean $\bar{c}_{\text{min-maj}}$ with per-pair std band; panel (b) plots the centroid cosine $c_{\text{centroid}} = \cos(\nabla L_1, \frac{1}{G-1}\sum_j \nabla L_j)$. Both metrics remain centered near zero over the 40-epoch horizon.

Gradient variance and DRO effectiveness. A natural hypothesis is that the near-zero cosines above are merely an artifact of high single-rollout REINFORCE variance: with a low-variance estimator, inter-group gradients might retain enough directional signal for DRO to exploit. Our POMO-100 experiment (Appendix C.6, Table 15) tests this directly. With $100\times$ variance reduction, a two-phase structure emerges: gradients are *aligned* ($\cos \approx +0.5$) during early training and become near-orthogonal ($\cos \approx -0.07$) after convergence. Critically, DRO cannot exploit either phase: in the aligned phase, ERM already follows the shared improvement direction; in the converged phase, group-specific gradients are nearly independent and re-weighting cannot redirect the optimization. DRO still matches ERM within 0.06% over 300 epochs, confirming that the ineffectiveness of loss re-weighting

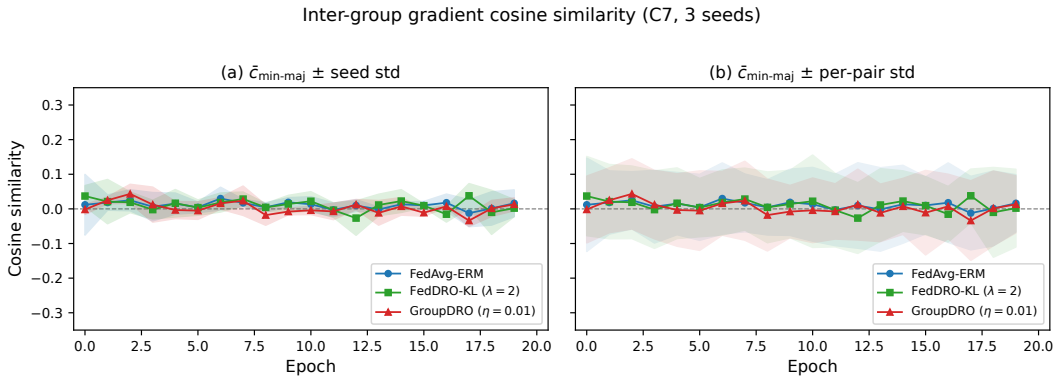

Figure 1: Inter-group gradient cosine similarity over epochs on condition C7 (cluster/grid, n=50, d=32, minority 1/16, 20 epochs, 3 seeds). **(a)** Pairwise mean $\bar{c}_{\text{min-maj}} \pm$ sample SD across seeds. **(b)** Pairwise mean $\pm$ per-pair std. The three-seed mean remains within ± 0.05 at every epoch.

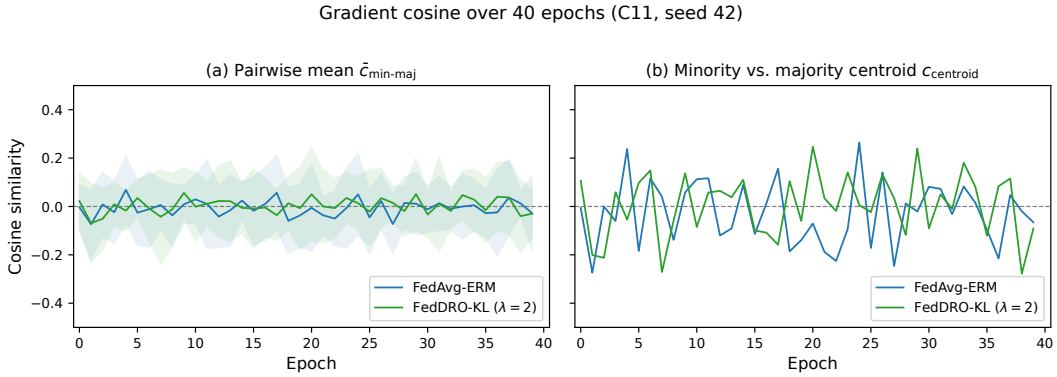

Figure 2: Gradient cosine similarity over 40 epochs on condition C11 (seed 42). **(a)** Pairwise mean $\bar{c}_{\text{min-maj}} \pm$ per-pair std. **(b)** Minority vs. majority centroid cosine c_{centroid}. Orthogonality persists through the full 40-epoch horizon for both ERM and FedDRO-KL.

is inherent to the inter-group gradient structure, not an artifact of estimator noise. The discrepancy with Jiang et al. [2022], who report DRO improvements in a centralized routing setting, remains an open question. Our centralized baseline (Table 13) also yields DRO≈ERM (worst-group difference -0.6%, p=0.71), ruling out federated aggregation as the cause. Possible explanations include differences in distribution pairs, problem scale, or their use of a CNN-based module for learning distribution-specific node patterns that may enable distribution-aware representation learning beyond simple loss re-weighting. An alternative approach is to directly optimize coherent risk measures such as CVaR via policy gradients [Tamar et al., 2015]; exploring such risk-aware objectives in the federated routing setting is another promising direction.

C Experimental conditions and full results

Table 5 lists all twelve experimental conditions. Table 6 reports worst-group cost for all conditions and methods.

Data generation. We pre-generate customer locations in $[0, 1]^2$ using the following procedures. These are our generator implementations and settings; they are not claimed to reproduce the generators of Jiang et al. [Jiang et al., 2022]. The six distribution types are:

- **Uniform:** sample each location independently and uniformly from $[0, 1]^2$.

- **Cluster:** sample two centers independently from $[0.2, 0.8]^2$. Assign each location uniformly to a center, sample it from a Gaussian with covariance $0.02^2 I$ around that center, and clip coordinates to $[0, 1]$.
- **Explosion:** start with uniform locations and sample a center uniformly from $[0, 1]^2$. Move each point within radius 0.45 radially outward to distance $0.45 + \gamma$, with independent $\gamma \sim \mathrm{Exp}(25)$, then jointly min–max normalize all coordinates to $[0, 1]$.
- **Implosion:** start with uniform locations; sample a center uniformly from $[0, 1]^2$ and one $s \sim \mathrm{Unif}(0, 1)$ per instance. Replace each point x within radius 0.45 by $c + s(x - c)$.
- **Expansion:** start with uniform locations and sample $b \sim \mathrm{Unif}(0, 1)$. For $b < 0.5$, sample $a \sim \mathrm{Unif}(0, 3)$; otherwise sample $a \sim \mathrm{Unif}(-3, 0)$. Move points within distance 0.15 of $y = ax + b$ perpendicular to the line, on the same side, to distance $0.15 + \gamma$, with independent $\gamma \sim \mathrm{Exp}(25)$. Then jointly min–max normalize all coordinates to $[0, 1]$.
- **Grid:** start with uniform locations and sample the lower left corner of an axis-aligned square uniformly from $[0, 0.85]^2$. If k points lie inside the square, replace them with the first k points, in row-major order, from an $m \times m$ grid including the square boundary, where $m = \lceil \sqrt{k} \rceil$.

For every instance, sample the depot uniformly from $[0, 1]^2$ and sample integer demands uniformly from $\{1, \ldots, 9\}$, then divide by vehicle capacity (40 for $n = 50$ and 50 for $n = 100$). For each client, we generate 12,800 training, 2,048 validation, and 1,024 test instances independently from its assigned distribution, using separate split seeds. Training data are stored in shards of 64 instances; at each round, each client samples a minibatch of B=128 by concatenating two randomly drawn shards (without replacement within an epoch). For generation seed s_0, the split seed is $s_0 + 10{,}000g + r$, where g is the zero-based code client index and $r = 1, 2, 3$ denotes training, validation, and test, respectively; instance index j adds $9{,}973j$ to that seed.

For the CVRP100 setting of Jiang et al., four clusters are used, whereas our $n = 100$ cluster generator uses two. Jiang et al. do not specify the remaining generator parameter values used here.

Six-distribution evaluation. The 16 clients are assigned by the paper's one-based indexing as follows: clients 1–10 use cluster, 11–12 grid, 13 explosion, 14 implosion, 15 expansion, and 16 uniform. Clients with the same distribution type form one of six evaluation groups; the worst-group cost is the maximum of their six group-level mean costs. The evaluation uses $n = 100$, $d = 32$, 40 epochs, 100 rounds per epoch, full participation, and $K = 1$. ERM and KL $\lambda = 5$ use 10 seeds; KL $\lambda = 2$, EMA $\lambda = 5$, and GroupDRO $\eta_q = 0.001$ use one seed each.

Table 5: Specification of the twelve experimental conditions. All conditions use G=16 clients and *cluster* as the majority distribution type. Sorted by minority type, n, d, minority ratio.

ID	Dist. pair	n	d	Minority ratio	Minority type
C1	cluster/explosion	50	32	$1/16$	explosion
C2	cluster/explosion	50	32	$4/16$	explosion
C3	cluster/explosion	50	32	$8/16$	explosion
C4	cluster/explosion	50	128	$1/16$	explosion
C5	cluster/explosion	100	32	$1/16$	explosion
C6	cluster/explosion	100	128	$1/16$	explosion
C7	cluster/grid	50	32	$1/16$	grid
C8	cluster/grid	50	32	$4/16$	grid
C9	cluster/grid	50	32	$8/16$	grid
C10	cluster/grid	50	128	$1/16$	grid
C11	cluster/grid	100	32	$1/16$	grid
C12	cluster/grid	100	128	$1/16$	grid

Conditions C7 (d=32) and C10 (d=128) share the cluster/grid pair, n=50, and minority ratio $1/16$, but use independently generated datasets: generation seeds 20260421 and 20260111, respectively.

Evaluation protocol. For the twelve main conditions, each checkpoint is evaluated by greedy decoding on all 16 clients' validation sets (2,048 instances per client). We report the maximum distribution-group mean cost at the final epoch (epoch 19 of 20); other evaluation horizons are specified in the corresponding captions.

Table 6: Worst-group cost (↓) across all twelve conditions. Final-epoch values (epoch 19 of 20), mean ± sample SD across 3 seeds. KL and EMA denote divisive implementations with nominal parameter λ. †Conditions also shown in Table 1.

		Ours (FedDRO-KL)		FedAvg-GroupDRO	
Condition	FedAvg-ERM	KL λ=2	EMA λ=5	η_q=0.01	η_q=0.001
C1† (expl, n=50, d=32, $\frac{1}{16}$)	$20.34_{\pm 1.51}$	$20.30_{\pm 1.61}$	$20.30_{\pm 1.66}$	$24.26_{\pm 0.66}$	$20.29_{\pm 1.65}$
C2 (expl, n=50, d=32, $\frac{4}{16}$)	$19.48_{\pm 1.53}$	$19.51_{\pm 1.48}$	$19.50_{\pm 1.48}$	$22.12_{\pm 1.04}$	$19.55_{\pm 1.46}$
C3 (expl, n=50, d=32, $\frac{8}{16}$)	$18.63_{\pm 1.40}$	$18.60_{\pm 1.42}$	$18.56_{\pm 1.35}$	$19.63_{\pm 1.51}$	$18.54_{\pm 1.41}$
C4 (expl, n=50, d=128, $\frac{1}{16}$)	$20.72_{\pm 2.78}$	$20.74_{\pm 2.87}$	$20.79_{\pm 2.86}$	$24.45_{\pm 4.49}$	$20.73_{\pm 2.85}$
C5† (expl, n=100, d=32, $\frac{1}{16}$)	$42.55_{\pm 2.97}$	$42.29_{\pm 2.88}$	$42.50_{\pm 2.94}$	$48.92_{\pm 3.04}$	$42.32_{\pm 3.46}$
C6 (expl, n=100, d=128, $\frac{1}{16}$)	$41.92_{\pm 7.22}$	$41.75_{\pm 7.06}$	$42.24_{\pm 7.09}$	$51.62_{\pm 10.84}$	$42.04_{\pm 7.23}$
C7 (grid, n=50, d=32, $\frac{1}{16}$)	$22.58_{\pm 1.98}$	$22.63_{\pm 2.07}$	$22.56_{\pm 2.05}$	$27.13_{\pm 1.30}$	$22.63_{\pm 1.94}$
C8† (grid, n=50, d=32, $\frac{4}{16}$)	$21.31_{\pm 2.00}$	$21.28_{\pm 2.00}$	$21.27_{\pm 2.00}$	$23.97_{\pm 2.17}$	$21.24_{\pm 2.01}$
C9 (grid, n=50, d=32, $\frac{8}{16}$)	$19.77_{\pm 2.05}$	$19.73_{\pm 2.14}$	$19.78_{\pm 2.02}$	$20.79_{\pm 2.12}$	$19.78_{\pm 2.19}$
C10 (grid, n=50, d=128, $\frac{1}{16}$)	$24.20_{\pm 3.54}$	$24.31_{\pm 3.67}$	$24.20_{\pm 3.55}$	$29.19_{\pm 6.41}$	$24.24_{\pm 3.75}$
C11† (grid, n=100, d=32, $\frac{1}{16}$)	$48.13_{\pm 0.91}$	$47.59_{\pm 1.01}$	$47.47_{\pm 1.11}$	$55.66_{\pm 0.54}$	$47.71_{\pm 0.83}$
C12 (grid, n=100, d=128, $\frac{1}{16}$)	$50.97_{\pm 5.35}$	$51.05_{\pm 5.57}$	$51.65_{\pm 5.00}$	$63.98_{\pm 11.16}$	$51.03_{\pm 5.56}$

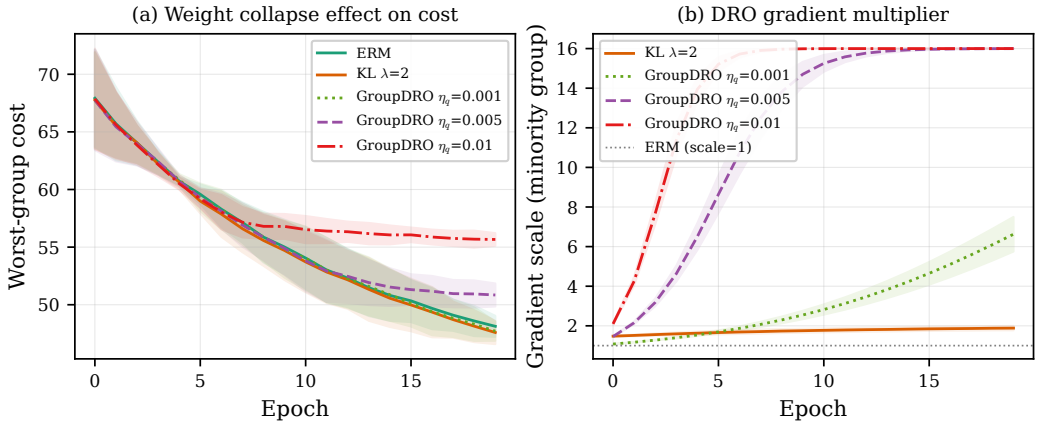

Figure 3: GroupDRO weight collapse on condition C11 (cluster/grid, n=100, minority $1/16$, 3 seeds, 20 epochs). Curves and shaded bands show the across-seed mean and sample SD at each epoch. **(a)** Worst-group cost: GroupDRO η_q=0.01 causes rapid divergence, while η_q=0.005 also leads to significant degradation. η_q=0.001 and KL λ=2 track ERM closely. **(b)** DRO gradient scale for the minority group: η_q=0.01 exhibits catastrophic concentration within the first 5 epochs. η_q=0.005 shows a slower but still unbounded growth. KL regularization maintains a stable ≈2× up-weighting throughout.

Gap to HGS references. For each held-out instance, we compute the relative gap $100(C_{\text{policy}} - C_{\text{HGS}})/C_{\text{HGS}}$; the mean gap averages over instances, and the worst-group gap is the maximum of the distribution-specific mean gaps. We compute the reference costs with MIT-licensed HGS-CVRP [Vidal, 2022, Vidal et al., 2012], invoked through the MIT-licensed PyHygese wrapper (v0.1.0)[2], which builds HGS-CVRP commit `1a92795`; the time limit is 10 CPU seconds.

[2] `https://github.com/chkwon/PyHygese`

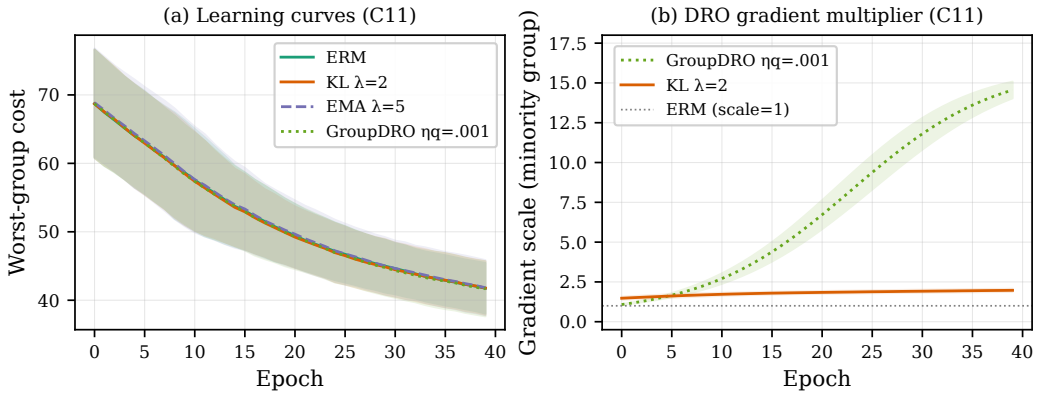

Figure 4: Training dynamics on condition C11 (cluster/grid, n=100, d=32, minority $1/16$, 40 epochs, 10 seeds). Curves and shaded bands show the mean and $\pm$ sample SD across seeds. **(a)** Worst-group cost learning curves. All DRO variants converge to similar costs as ERM, with no statistically significant improvement. **(b)** DRO gradient scale (effective gradient multiplier) applied to the minority group. At epoch 40, GroupDRO η_q=0.001 places 91% of total weight on the minority client on average (max/min weight ratio $\approx$177$\times$), with a minority-client gradient scale of $\approx$15$\times$. For KL λ=2, the mean max/min ratio is $\approx$2.1$\times$ at the final epoch and ranges from 1.53 to 2.12 across epochs. Despite significant gradient re-weighting by both methods, worst-group performance remains similar.

Table 7: Gaps to HGS-CVRP reference costs (10 CPU-s time limit). Entries are mean $\pm$ sample SD across seeds, except B-6a (one run, seed 42). HGS is a heuristic and does not certify optimality.

Setting	Method	Seeds	Mean gap (%)	Worst-group gap (%)
C1, 20 ep.	ERM	3	79.3 ± 17.1	117.5 ± 15.5 (expl.)
	KL λ=2	3	79.7 ± 17.0	117.2 ± 16.6 (expl.)
C5, 20 ep.	ERM	3	167.1 ± 31.2	210.7 ± 20.4 (expl.)
	KL λ=2	3	166.7 ± 27.9	209.0 ± 19.9 (expl.)
C8, 20 ep.	ERM	3	92.5 ± 20.0	107.5 ± 18.6 (grid)
	KL λ=2	3	91.8 ± 19.7	106.9 ± 18.9 (grid)
C11, 20 ep.	ERM	3	167.9 ± 23.6	209.6 ± 5.1 (grid)
	KL λ=2	3	167.6 ± 25.5	206.3 ± 5.3 (grid)
C11 spec., 40 ep.	ERM	10	127.8 ± 26.1	170.2 ± 24.0 (grid)
	KL λ=2	10	127.0 ± 27.5	170.4 ± 25.4 (grid)
B-6a, 300 ep.	ERM	1	36.8	38.4 (uniform)
	KL λ=5	1	36.8	38.5 (uniform)

C.1 Temperature sensitivity (λ ablation)

Table 8 reports the effect of the KL temperature λ on worst-group cost for the four representative conditions from Table 1. All three values produce results statistically indistinguishable from ERM. As λ increases, the softmax weights in Eq. (4) flatten toward the uniform prior, so λ=10 behaves nearly identically to ERM—consistent with the theoretical prediction that large λ recovers uniform weighting. The most active re-weighting occurs at λ=2, which produces the largest (though still non-significant) deviations from ERM. Based on this sweep, we select λ=2 for FedDRO-KL and λ=5 for FedDRO-KL-EMA in the main experiments.

C.1.1 Extended λ sweep: from collapse to ERM

To characterise the full sensitivity landscape, we extend the temperature sweep to seven values $\lambda \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$ on condition C1 with 3 seeds. Table 9 and Figures 5–6 report the results.

Table 8: Temperature ablation: final-epoch worst-group cost (epoch 19 of 20; ↓) for FedDRO-KL across $\lambda \in \{2, 5, 10\}$. Mean ± sample SD across 3 seeds. ERM is shown for reference. All differences from ERM are within seed-level noise ($p > 0.05$).

Condition	FedAvg-ERM	KL λ=2	KL λ=5	KL λ=10
C1 (expl, n=50, $\frac{1}{16}$)	$20.34_{\pm 1.51}$	$20.30_{\pm 1.61}$	$20.27_{\pm 1.62}$	$20.32_{\pm 1.57}$
C5 (expl, n=100, $\frac{1}{16}$)	$42.55_{\pm 2.97}$	$42.29_{\pm 2.88}$	$42.54_{\pm 3.38}$	$42.34_{\pm 3.25}$
C8 (grid, n=50, $\frac{4}{16}$)	$21.31_{\pm 2.00}$	$21.28_{\pm 2.00}$	$21.18_{\pm 2.08}$	$21.23_{\pm 2.05}$
C11 (grid, n=100, $\frac{1}{16}$)	$48.13_{\pm 0.91}$	$47.59_{\pm 1.01}$	$47.90_{\pm 0.60}$	$47.85_{\pm 0.78}$

Normalised DRO weights. The optimal KL-DRO weights ξ_g in Eq. (4) satisfy $\xi_g \propto p_g \exp(L_g/\lambda)$, where L_g is the (EMA-smoothed) group loss. We write $\xi_{\min}$ for the weight on the minority group; it controls how much gradient signal comes from that group. Table 9 shows that $\xi_{\min}$ varies dramatically: at λ=0.1, the minority receives $\xi_{\min} \approx 99.9\%$ of the total weight—effectively training on a single 64-sample client—while at λ=10, $\xi_{\min} \approx 6.9\%$, close to the uniform $1/G = 6.25\%$.

Three regimes. The results reveal three operating regimes (Figure 5):

1. **Collapse regime ($\lambda \leq 0.1$):** The KL penalty is too weak to regularise the weights. The normalised weight $\xi_{\min}$ saturates near 1.0, concentrating virtually all gradient signal on the minority client. This mirrors GroupDRO collapse (Appendix B.1) in a "soft" form: rather than a hard argmax on the worst group, the softmax temperature produces an extreme distribution. Worst-group cost degrades by $+14.6\%$ relative to ERM (Mann–Whitney U test, $p = 0.0015$, comparing the three λ=0.1 runs with the 18 runs at $\lambda \geq 0.2$).
2. **Insensitive regime ($0.2 \leq \lambda \leq 10$):** Despite large variation in $\xi_{\min}$ (from 93.9% to 6.9%), worst-group cost remains within $\pm 0.4\%$ of ERM across this $50\times$ range of λ. This insensitivity is explained by inter-group gradient near-orthogonality (Section 6.2): when group gradients are nearly orthogonal, re-weighting merely rescales independent gradient components without redirecting the overall update. One-way ANOVA across all seven λ values (including λ=0.1) yields $F = 1.75$, $p = 0.18$; excluding λ=0.1, $F < 0.01$, $p > 0.99$.
3. **ERM recovery ($\lambda \geq 10$):** Large λ strengthens the KL penalty, pushing the weights toward the uniform prior p. At λ=10, $\xi_{\min}$=6.9% is already within 0.7 pp of uniform ($1/G = 6.25\%$).

Practical guidance. In this sweep, the tested settings with $\lambda \geq 0.2$ yield ERM-comparable performance. We recommend $\lambda \in [1, 5]$ as a default range, balancing DRO expressiveness against proximity to ERM.

Table 9: Extended temperature sweep on C1: worst-group cost and normalised minority weight $\xi_{\min}$ at epoch 19. Mean ± sample SD over 3 seeds. Δworst is relative to ERM.

λ	Avg. cost	Worst cost	Δworst (%)	$\xi_{\min}$
0.1	$14.21_{\pm 0.58}$	$23.30_{\pm 0.41}$	$+14.6$	0.999
0.2	$11.34_{\pm 0.92}$	$20.27_{\pm 1.65}$	-0.3	0.939
0.5	$11.38_{\pm 0.95}$	$20.39_{\pm 1.60}$	$+0.3$	0.372
1	$11.33_{\pm 0.93}$	$20.29_{\pm 1.52}$	-0.2	0.166
2	$11.36_{\pm 0.95}$	$20.30_{\pm 1.61}$	-0.2	0.103
5	$11.34_{\pm 0.92}$	$20.27_{\pm 1.62}$	-0.3	0.077
10	$11.38_{\pm 0.94}$	$20.32_{\pm 1.57}$	-0.1	0.069
ERM	$11.36_{\pm 0.97}$	$20.34_{\pm 1.51}$	—	0.0625

C.2 Local-step ablation (K ablation)

We study the effect of the number of local SGD steps K per communication round on condition C1 (cluster/explosion, n=50, d=32, minority $1/16$, G=16). The total number of communication rounds is fixed at T=2,000, so increasing K linearly increases the total number of gradient steps while keeping the communication budget constant. Table 10 reports the results.

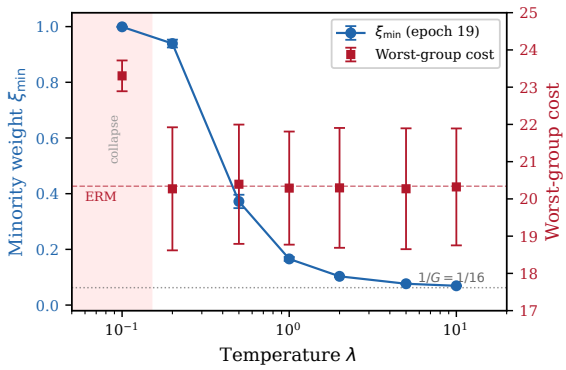

Figure 5: Temperature sensitivity on C1 at the final epoch (epoch 19 of 20). **Left axis** (blue circles): normalised minority weight $\xi_{\min}$. **Right axis** (red squares): worst-group cost. Markers and error bars show the mean and sample SD across 3 seeds. The shaded region marks the collapse regime ($\lambda \leq 0.1$), where $\xi_{\min} \to 1$ and cost rises sharply. For $\lambda \geq 0.2$, worst-group cost is flat and indistinguishable from ERM (dashed line).

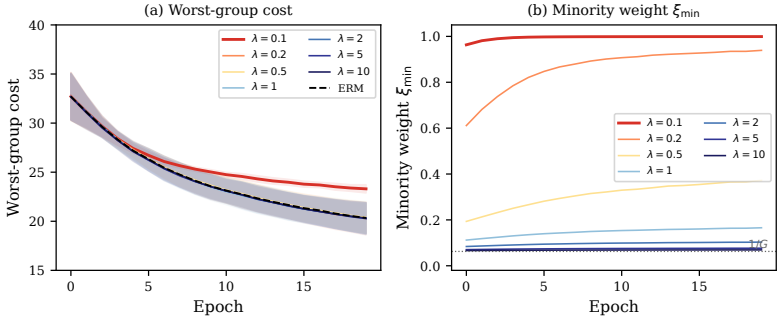

Figure 6: Training dynamics across λ values on C1. Lines show across-seed means; shaded bands in panel (a) show $\pm$ sample SD across 3 seeds at each epoch. **(a)** Worst-group cost vs. epoch. λ=0.1 (red) converges to a distinctly higher plateau; all other λ overlap with ERM. **(b)** Normalised minority weight $\xi_{\min}$ vs. epoch. λ=0.1 rapidly saturates at $\xi_{\min} \approx 1$, while $\lambda \geq 1$ remains near $1/G$.

Increasing K from 1 to 3 reduces worst-group cost by 23% (from 20.34 to 15.72 for ERM), confirming that additional local computation substantially improves solution quality in this federated routing setting. The improvement from K=2 to K=3 (15.72 vs. 17.20, -8.6%) is smaller than from K=1 to K=2 (17.20 vs. 20.34, -15.4%), exhibiting diminishing returns. This is consistent with the $\mathcal{O}(K^2H^2)$ client-drift term in Theorem 1: larger K increases local drift, but the additional gradient steps more than compensate within this range. Critically, the gap between FedDRO-KL and ERM remains within 0.5% at all tested values of K.

C.3 Partial participation ablation ($|S|$ ablation)

This section provides the full table and additional analysis for the partial participation ablation. We evaluate the effect of partial client participation by varying $|S^{(t)}|$, the fixed number of clients sampled uniformly without replacement per round, from 16 (full participation) down to 1 (6.25%) on condition C1. Table 11 reports the results.

The observed worst-group means range from 20.00 to 20.40 without a monotone trend over the tested subset sizes. The seed-level sample SD of worst-group costs ranges from 1.22 to 1.86. Across the high-participation settings $|S| \in \{12, 13, 14, 15\}$, the largest absolute relative ERM–KL difference in worst-group mean cost is 0.62% at $|S| = 13$ (below 0.7%). These three-seed measurements do not isolate the effect of client-sampling variance.

Table 10: Final-epoch local-step ablation (epoch 19 of 20) on C1: worst-group and average cost (↓) for varying K. Mean ± sample SD across 3 seeds. Increasing K improves performance (more gradient steps per round), while FedDRO-KL λ=2 remains statistically indistinguishable from ERM at every K.

	Average cost		Worst-group cost	
K	FedAvg-ERM	KL λ=2	FedAvg-ERM	KL λ=2
1	$11.36_{\pm 0.97}$	$11.36_{\pm 0.95}$	$20.34_{\pm 1.51}$	$20.30_{\pm 1.61}$
2	$9.08_{\pm 0.54}$	$9.11_{\pm 0.58}$	$17.20_{\pm 1.97}$	$17.17_{\pm 2.07}$
3	$8.37_{\pm 0.35}$	$8.35_{\pm 0.34}$	$15.72_{\pm 1.38}$	$15.65_{\pm 1.52}$

Table 11: Partial participation ablation on C1 (cluster/explosion, n=50, d=32, G=16, minority $1/16$; 20 epochs, 3 seeds). Entries are final-epoch mean ± sample SD. Each round samples a fixed subset uniformly without replacement; $|S|$=1 corresponds to 6.25% participation.

	Average cost		Worst-group cost	
$\lvert S\rvert$	FedAvg-ERM	KL λ=2	FedAvg-ERM	KL λ=2
16	$11.36_{\pm 0.97}$	$11.36_{\pm 0.95}$	$20.34_{\pm 1.51}$	$20.30_{\pm 1.61}$
15	$11.33_{\pm 0.94}$	$11.36_{\pm 1.04}$	$20.25_{\pm 1.58}$	$20.34_{\pm 1.70}$
14	$11.38_{\pm 0.90}$	$11.34_{\pm 0.90}$	$20.38_{\pm 1.55}$	$20.34_{\pm 1.46}$
13	$11.29_{\pm 0.92}$	$11.32_{\pm 0.89}$	$20.26_{\pm 1.68}$	$20.38_{\pm 1.53}$
12	$11.36_{\pm 0.99}$	$11.32_{\pm 0.94}$	$20.36_{\pm 1.59}$	$20.33_{\pm 1.51}$
8	$11.32_{\pm 0.92}$	$11.36_{\pm 1.00}$	$20.36_{\pm 1.54}$	$20.35_{\pm 1.40}$
4	$11.32_{\pm 0.99}$	$11.41_{\pm 0.92}$	$20.00_{\pm 1.46}$	$20.21_{\pm 1.22}$
1	$11.41_{\pm 0.89}$	$11.44_{\pm 0.85}$	$20.30_{\pm 1.37}$	$20.40_{\pm 1.86}$

C.4 Independent learning baseline

This section provides the full table and additional details for the independent learning comparison discussed in Section 6.2. To isolate the benefit of federated training, we compare against *independent learning*, where each client trains a model using only its own local data without any communication. We train independent models for two representative clients on condition C1: client 0 (minority, explosion distribution) and client 1 (majority, cluster distribution). All models are then evaluated on the validation data of *all* 16 clients, ensuring a fair comparison on an identical evaluation set. Table 12 reports the results.

Table 12: Independent learning vs. federated training on C1. All models evaluated on all 16 clients' validation data. Final-epoch values (epoch 19), mean ± sample SD across 3 seeds. The independent minority-client model achieves lower cost on its own distribution but catastrophically degrades on unseen distributions, while the federated model generalizes across all groups.

Method	Avg. cost	cluster	explosion
FedAvg-ERM	$11.36_{\pm 0.97}$	$10.76_{\pm 0.95}$	$20.34_{\pm 1.51}$
FedDRO-KL λ=2	$11.36_{\pm 0.95}$	$10.77_{\pm 0.92}$	$20.30_{\pm 1.61}$
Indep. client 1 (majority/cluster)	$11.34_{\pm 0.85}$	$10.72_{\pm 0.82}$	$20.68_{\pm 1.38}$
Indep. client 0 (minority/explosion)	$14.34_{\pm 3.29}$	$14.15_{\pm 3.40}$	$17.21_{\pm 1.87}$

The independent minority-client model (client 0) achieves 15.4% lower cost on its own explosion distribution (17.21 vs. 20.34), but degrades by 31.5% on the cluster distribution (14.15 vs. 10.76) and 26.3% overall (14.34 vs. 11.36), demonstrating a failure to generalize beyond its local data. The independent majority-client model (client 1) performs comparably to federated training ($\Delta < 2\%$ on all metrics), as expected: the majority distribution accounts for $15/16$ of clients, providing sufficient diversity even in isolation.

These results highlight the value of federated learning for heterogeneous routing deployments: it enables the minority client to benefit from the majority's data diversity while maintaining competitive performance on its own distribution, a trade-off that independent training cannot achieve.

C.5 Centralized learning baseline

To test whether the observed DRO≈ERM similarity is caused by federated aggregation noise or by the problem itself, we run a *centralized* learning experiment on condition C1 with $G{=}2$ clients holding true group labels: client 0 (explosion, minority) and client 1 (cluster, majority), each receiving equal weight ($1/G = 0.5$). With $K{=}1$ local step and full participation, FedAvg reduces to centralized SGD, so any DRO effect should be maximally visible. Table 13 reports the results.

Table 13: Centralized learning on C1 ($G{=}2$, true group labels). Final-epoch values (epoch 19), mean $\pm$ sample SD across 3 seeds. GroupDRO weights are reported as $w{=}(1, 0)$ at the displayed precision even in the centralized setting.

Method	Avg. cost	cluster	explosion (worst)
Central ERM	$15.41_{\pm 1.82}$	$12.27_{\pm 2.44}$	$18.56_{\pm 1.42}$
Central KL $\lambda{=}2$	$15.37_{\pm 1.88}$	$12.29_{\pm 2.42}$	$18.46_{\pm 1.46}$
Central GroupDRO $\eta{=}0.01$	$16.46_{\pm 1.93}$	$13.75_{\pm 2.76}$	$19.17_{\pm 1.52}$
FedAvg-ERM ($G{=}16$)	$11.36_{\pm 0.97}$	$10.76_{\pm 0.95}$	$20.34_{\pm 1.51}$
FedDRO-KL ($G{=}16$, $\lambda{=}2$)	$11.36_{\pm 0.95}$	$10.77_{\pm 0.92}$	$20.30_{\pm 1.61}$

FedDRO-KL achieves only -0.6% worst-group cost change relative to ERM ($p = 0.71$), confirming that similar DRO–ERM performance persists even without federated aggregation. GroupDRO collapses to weights $(1.0, 0.0)$ across all three seeds at the displayed precision, demonstrating that weight collapse also occurs without federated aggregation.

On group-specific final-epoch validation means, centralized ERM has higher cluster cost than FedAvg-ERM (12.27 vs. 10.76; $+14.1\%$) and lower explosion cost (18.56 vs. 20.34; -8.8%), relative to the corresponding federated group means. The overall average-cost columns are not directly comparable: the centralized run weights its two groups equally, whereas the federated run has a 15:1 client ratio. It also uses one cluster client's training data rather than the federated run's 15 cluster clients.

C.6 POMO multi-rollout experiment

A natural hypothesis is that the DRO–ERM similarity observed above stems from the high variance of single-rollout REINFORCE: with a low-variance estimator, inter-group gradients might retain enough directional signal for DRO re-weighting to improve worst-group performance. To test this, we run federated experiments using POMO [Kwon et al., 2020] with $P{=}100$ multi-start rollouts—a variance reduction factor of $100\times$ relative to single-rollout REINFORCE. We adopt a two-group setting ($G{=}2$: uniform and explosion distributions, each holding one client) with $d{=}128$, $\mathrm{lr}{=}10^{-4}$, SGD with momentum, and 300 epochs (100 steps per epoch); this is not a reproduction of the setup in Jiang et al. [2022]. Our uniform/explosion mixture is 1:1 and uses the generators above, whereas their typical-to-atypical mixture is 10:1 and their POMO model uses an additional CNN-based module for distribution-aware features.

Table 14: POMO-100 multi-rollout experiment at the final epoch (epoch 299; one run, seed 42; $G{=}2$, uniform/explosion, $d{=}128$, 300 epochs). Worst group is uniform in all cases.

Method	Avg. cost	uniform	explosion	worst (uniform)
FedAvg-ERM	19.999	21.438	18.561	21.438
FedDRO-KL ($\lambda{=}5$)	20.005	21.452	18.558	21.452
FedAvg-GroupDRO ($\eta_q{=}0.01$)	21.064	21.970	20.157	21.970

Table 14 shows the results for a single seed. FedDRO-KL matches ERM within 0.06% worst-group cost ($\Delta{=}+0.01$ in absolute terms), confirming that even with $100\times$ variance reduction, DRO re-weighting does not improve over ERM. GroupDRO collapses to weights $(1.0, 0.0)$ at the

displayed precision by epoch 31 and degrades worst-group cost by $+2.5\%$ (21.97 vs. 21.44); the milder degradation compared to the G=16 setting ($+5$–26% in Table 6) suggests that POMO's lower gradient variance partially mitigates the damage from weight collapse, but does not prevent it.

The inter-group gradient cosine similarity reveals a two-phase structure that is hidden under single-rollout REINFORCE noise. Table 15 reports $\cos(\nabla L_0, \nabla L_1)$ averaged over 50-epoch windows for both ERM and FedDRO-KL.

Table 15: Inter-group gradient cosine by training phase (POMO-100, G=2, seed 42). Mean $\pm$ sample SD computed over epochs within each window.

Epoch range	ERM	FedDRO-KL (λ=5)
0–49	$+0.474 \pm 0.130$	$+0.496 \pm 0.116$
50–99	$+0.452 \pm 0.099$	$+0.420 \pm 0.103$
100–149	$+0.153 \pm 0.134$	$+0.136 \pm 0.164$
150–199	-0.089 ± 0.163	-0.017 ± 0.148
200–249	-0.071 ± 0.208	-0.072 ± 0.186
250–299	-0.071 ± 0.178	-0.090 ± 0.187
All 300 ep	$+0.142 \pm 0.287$	$+0.146 \pm 0.279$

Two observations emerge, visualised in Figure 7:

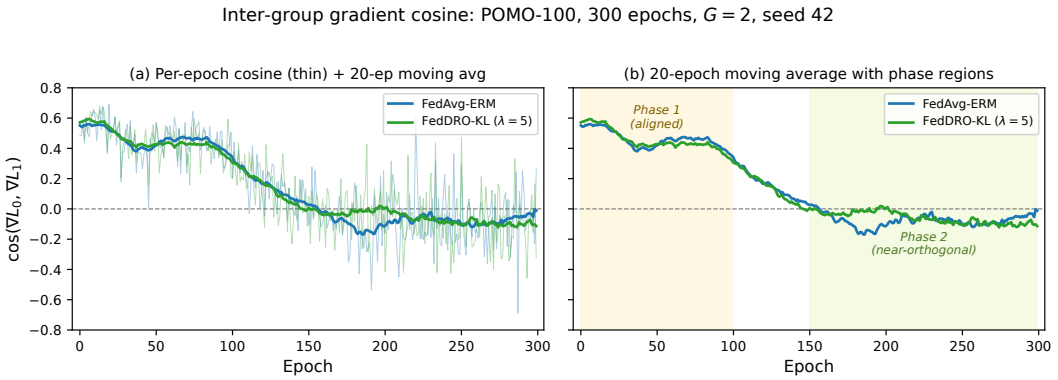

Figure 7: Inter-group gradient cosine $\cos(\nabla L_0, \nabla L_1)$ over 300 epochs with POMO-100 (G=2, seed 42). **(a)** Per-epoch values (thin lines) with 20-epoch moving average (thick lines). **(b)** Moving average with phase regions highlighted. Phase 1 (epochs 0–99): gradients are strongly aligned ($\cos \approx +0.5$). Phase 2 (epochs 150–299): gradients settle to near-orthogonal ($\cos \approx -0.07$). DRO re-weighting is ineffective in both phases.

Phase 1 (epochs 0–99): aligned gradients. Both groups share a strong common improvement direction ($\cos \approx +0.5$). Because the gradients are aligned, re-weighting is equivalent to a per-group learning-rate adjustment—it changes the *magnitude* of the update along an already shared direction but cannot redirect it. ERM already optimises this shared direction efficiently, so DRO up-weighting of the minority provides no additional benefit.

Phase 2 (epochs 150–299): near-orthogonal with weak conflict. As the model converges, the shared improvement direction is exhausted and the cosine settles at $\cos \approx -0.07$ (both ERM and KL, sample SD across epochs ≈ 0.18). This weak negative correlation implies that a small fraction of the gradient energy is conflicting, but $\sin(\arccos(-0.07)) \approx 0.998$, meaning 99.8% of the gradient information lies in orthogonal dimensions. Re-weighting therefore has negligible capacity to redirect the optimization, and the residual conflict is too weak to exploit. Moreover, gradient magnitudes are small near convergence, further diminishing the impact of any re-scaling.

Comparison with single-rollout REINFORCE. In the C11-specification, 40-epoch seed-42 runs reported in Table 4, the epoch-averaged pairwise minority–majority cosine is -0.007 for ERM and $+0.003$ for KL; the maximum absolute per-epoch values are 0.073 and 0.069, respectively. Thus

$|\bar{c}| \leq 0.01$ for the epoch means and $|\bar{c}^{(t)}| < 0.08$ for every epoch in these two runs. These bounds do not describe Figure 1, which uses C7 data and three seeds. Compared with this C11-specification run, the POMO-100 experiment exhibits a two-phase gradient pattern: once variance is reduced, early alignment becomes visible, but this alignment is one that ERM already exploits. The late-training near-orthogonality is consistent across both estimators, confirming that it reflects the inter-group gradient structure rather than estimator noise.

In summary, the low-variance estimator does not restore a directional gradient signal that DRO could exploit. In Phase 1, the signal exists but favors all groups equally; in Phase 2, group-specific signals are nearly independent. DRO re-weighting is ineffective in both regimes, consistent with the similar DRO–ERM performance observed in these experiments.

Final-epoch KL weights. The divisive FedDRO-KL implementation has DRO scales of $(1.013, 0.987)$ at the final epoch in this 300-epoch experiment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state four contributions—framework (Section 4), convergence guarantees (Section 5), weight-collapse analysis (Section 3), and empirical insights (Section 6)—each supported by the corresponding theoretical and experimental results. The Conclusion (Section 7) synthesizes these as a robustness certificate and empirical ERM-level quality, and discusses the observed gradient orthogonality in CVRP.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion (Section 7) contains a dedicated "Limitations" paragraph discussing: synthetic-only CVRP data, limited scale ($n \leq 100$, G=16), partial participation tested only on C1 down to $|S|$=1, the sampling and estimator assumptions, full-participation $K = 1$ EMA theory, missing CVaR/adaptive step-size extensions, and untested alternative heterogeneity types (demand/capacity) and real-world instances. Section 6.2 additionally reports that no DRO variant achieves statistically significant improvement over ERM (10 seeds on C11, all $p > 0.28$). Appendix A.3 details the differences between theory and implementation, including divisive weighting, the baseline, BatchNorm, clipping, momentum, and weight decay.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are stated formally in Sections 3 and 5 and summarized in Appendix A.4. Theorems, propositions, and corollaries explicitly reference the required assumptions. Complete proofs are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6.1 specifies the architecture (Attention Model with 3 encoder layers), the main and supplementary settings covering six customer-location distribution types, hyperparameter sweeps (λ, β values), number of rounds, and random seeds. Appendix C details the generator procedures, parameters, client assignments, data splits, and six-distribution evaluation. Pseudocode is provided in Algorithms 1–2. Full optimizer details (SGD with momentum 0.9, weight decay 10^{-4}, batch size $B{=}128$) and runtime are in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is proprietary to the authors' institution and cannot be released at this time. However, all algorithmic details, hyperparameters, data generation procedures, and pseudocode are described in sufficient detail for independent reimplementation (Section 6.1, Algorithms 1–2, and the Appendix).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6.1 specifies the learning rate (10^{-3}), gradient clipping (1.0), KL temperature sweep ($\lambda \in \{2, 5, 10\}$; extended in the ablation), EMA parameters (β_u=β_v=0.05, β_d=0.1), training duration (2000 rounds), and random seeds. Optimizer details (SGD, momentum 0.9, weight decay 10^{-4}), batch size (B=128), hardware (single NVIDIA V100 32GB), and full condition specifications are in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports mean $\pm$ sample standard deviation over 3 random seeds for representative conditions; full results for all twelve conditions appear in Table 6 (Appendix). A supplementary 10-seed analysis on condition C11 (Section 6.2) provides paired t-test p-values, with paired 95% confidence intervals and Cohen's d in Remark A.3. These tests detect no statistically significant improvement or degradation relative to ERM (all $p > 0.28$); non-significance is not interpreted as statistical equivalence. The standard deviations capture variability over random initialization seeds. The additive-centering ablation reports sample SD across three seed-level last-five-epoch cost averages; single-seed reference rows have no uncertainty estimate (Table 3).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A states all training experiments were conducted on a single NVIDIA V100 (32GB). Training time per run is reported ($\sim$10.5 hours for n=100, d=32; $\sim$5 hours for n=50). Communication cost per round is quantified ($\sim$119K parameters for d=32; $\sim$693K for d=128). The HGS reference evaluation covered 135,168 instances at a 10-second CPU-time limit (approximately 375 CPU-hours), run with 20-way parallelism on a 28-core Xeon E5-2680 v4 Linux host for approximately 20.4 wall-clock hours. This CPU solver time is separate from GPU training; a combined GPU-hour total across all reported and exploratory runs is not available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms to it in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work addresses foundational optimization methodology for vehicle routing and does not study a deployed societal-impact application. We do not identify direct societal impacts of the work; the federated setting does not by itself provide a formal privacy guarantee.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release pre-trained models or datasets that pose risks for misuse. The trained CVRP routing policies are domain-specific and have no dual-use potential.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We build upon the Attention Model architecture of Kool et al. (2019), which is properly cited. Our implementation is original code. All referenced prior works are properly cited. HGS-CVRP reference solutions are computed with the MIT-licensed solver (Vidal, 2022; Vidal et al., 2012) through PyHygese v0.1.0, also MIT-licensed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not release new datasets, models, or code assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methods in this research do not involve LLMs as any important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.