FINE-TUNING GENERATIVE MODELS FOR EXTREME EVENTS VIA CVAR-PENALIZED WASSERSTEIN GRADIENT FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

In many high-stakes domains, extreme events carry substantial consequences, yet learning the heavy-tailed distributions that govern them from finite samples remains challenging: the quantities of interest are driven by a few extreme observations, so the tail region is under-sampled relative to its importance. Even generative models tailored for heavy tails capture the tail region inadequately in practice. We propose the Conditional Value-at-Risk (CVaR)-penalized Generative Particle Algorithm (CVaR-GPA), a tail-agnostic algorithm for fine-tuning generative models toward heavy-tailed targets, built as a time discretization of the Wasserstein gradient flow of the Lipschitz-regularized KL divergence penalized by a CVaR discrepancy term. Such a flow can be initialized from the output samples of the pre-trained model, which are then transported along the gradient descent of the loss functional, without requiring access to the pre-trained model's internal architecture. The Lipschitz-regularized KL divergence requires minimal assumptions on the target, while the CVaR penalty focuses the flow on the tail. The CVaR penalty depends on the target only through a scalar tail statistic, inducing a velocity field that remains active in the under-sampled tail region at a dimension-free estimation cost. The resulting velocity field, and hence the depth of the transport map, implicitly adapts to the target distribution, without any target-specific modifications. Across four targets, including two real-world, high-dimensional datasets (daily streamflow in the Ohio River basin ($d = 64$) and the Fama–French portfolios ($d = 25$) with tail indices ranging from 1.05 to 3.34), fine-tuning with CVaR-GPA reduces global and tail errors by geometric-mean factors of $14.0\times$ and $9.8\times$, respectively, across seven pre-trained models spanning GANs, diffusion models, and other generative flows, with a single set of hyperparameters.

1 INTRODUCTION

Heavy-tailed distributions arise in a wide range of application domains, including finance and insurance (Albrecher & Asmussen, 2006; Embrechts & Veraverbeke, 1982; Grossi & Kunreuther, 2005), climate science (Farazmand & Sapsis, 2019; Merz et al., 2022), and biology (Cirillo & Taleb, 2020; Lynn et al., 2024), where accurate simulations of extreme events are essential. Under heavy tails, the quantities of interest are driven by a few extreme observations (Embrechts et al., 1997; Foss et al., 2013; Nair et al., 2022): tail events are rare but not negligible. This makes learning such distributions from finite samples statistically challenging: only a small fraction of the samples falls in the tail region, which is therefore under-sampled relative to its importance. Standard objectives for learning distributions weigh all samples equally and are thus dominated by the bulk of the distribution, making them ill-suited for heavy-tailed targets. Despite their ability to simulate complex target distributions, standard generative models, which typically consist of globally Lipschitz continuous transport maps and light-tailed source distributions, necessarily output a light-tailed distribution (Jaini et al., 2020), and thus cannot capture heavy-tailed targets accurately.

The necessity and the challenges of learning heavy-tailed targets have motivated a growing class of generative models designed specifically for heavy-tailed targets, including extreme-value GANs (Allouche et al., 2022; 2026; Bhatia et al., 2021), Pareto GAN (Huster et al., 2021), Tail-GAN (Cont et al., 2025), and generative flows tailored for heavy-tailed targets (Hickling & Prangle, 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Laszkiewicz et al., 2022; Pandey et al., 2025). However, the statistical challenges inherent to learning heavy-tailed targets persist in these models, which capture the tails inadequately in practice. Models with fixed, target-independent architectural depths (Allouche et al., 2022; 2026; Cont et al., 2025) may not be deep enough to reach the extreme regions of the target distribution. Although the depth of the model proposed in Gu et al. (2024) adapts to the target by construction, due to the scarcity of samples in the tail region, its training terminates in practice before the extreme regions are adequately captured. Moreover, many models tailored for heavy-tailed targets require prior knowledge or accurate estimation of the target tail-decay rates (Huster et al., 2021; Hickling & Prangle, 2025; Laszkiewicz et al., 2022; Pandey et al., 2025), which itself is a challenging finite-sample statistical problem, again due to the under-sampling of the tail region. More recently, Chang & Sapsis (2026) propose an objective for capturing heavy-tailed behavior without requiring training samples from the tail. However, the method relies critically on the choice of the reference measure in the Wasserstein-1 term, which in turn requires some prior knowledge of the target.

We propose a fine-tuning flow and develop its corresponding particle algorithm, *CVaR-GPA*, the *Conditional Value-at-Risk (CVaR)-penalized Generative Particle Algorithm*, which mitigates these drawbacks to better learn heavy-tailed distributions. Rather than designing and training yet another architecture, we retain the capabilities of existing pre-trained models tailored for heavy-tailed distributions, and correct their output samples where necessary to overcome their limitations. Among generative frameworks, Wasserstein gradient flows (Jordan et al., 1998; Otto, 2001), which describe the gradient descent of a free energy functional $\mathcal{F}$ on the space of probability measures, are particularly suitable for fine-tuning. Since the velocity field of a Wasserstein gradient flow is determined by the variational derivative of $\mathcal{F}$, desired properties can be embedded into the flow through the choice of $\mathcal{F}$. To fine-tune existing models toward heavy-tailed targets without accessing or modifying their internal architecture, we require the fine-tuning flow to have the following properties.

- **(I)** The fine-tuning flow can be initialized from any empirical distribution supported on finite samples, in particular from the output samples of a pre-trained model. Thus, fine-tuning requires neither access to the pre-trained model's internal architecture nor any modifications specific to the pre-trained model.
- **(II)** The velocity field of the fine-tuning flow places additional weight on the tail region and remains active there despite its under-sampling, even for high-dimensional targets.
- **(III)** The velocity field of the fine-tuning flow is target-adaptive, and hence so is the terminal time of the flow, which determines the depth of the fine-tuning transport map. Thus, the fine-tuning flow self-regulates across targets without requiring target-specific modifications.

Let P^{pre} and P^{tar} denote the output distribution of a pre-trained model and the target distribution. We design $\mathcal{F}$ as the Lipschitz-regularized KL divergence (Dupuis & Mao, 2022) penalized by a smoothed Conditional Value-at-Risk (CVaR) discrepancy term (Kouri & Surowiec, 2016) given by

$$\mathcal{F}(Q; P^{\mathrm{tar}}) = D^L_{\mathrm{KL}}(Q \| P^{\mathrm{tar}}) + \lambda \left(\Delta C_h(Q; P^{\mathrm{tar}})\right)^2, \tag{1}$$

where $\Delta C_h(Q; P^{\mathrm{tar}}) := \mathrm{CVaR}^{P^{\mathrm{tar}},g}_{\alpha,h} - \mathrm{CVaR}^{Q,g}_{\alpha,h}$ denotes the smoothed CVaR discrepancy with smoothing parameter $h > 0$ of $g(X)$ between $X \sim Q$ and $X \sim P^{\mathrm{tar}}$, at level $\alpha \in (0,1)$, for a given risk function $g : \mathbb{R}^d \to [0,\infty)$, D^L_{KL} denotes the Lipschitz-regularized KL divergence with Lipschitz constant L, and $\lambda > 0$ is a hyperparameter controlling the relative weight of the CVaR penalization. A divergence term is needed in $\mathcal{F}$ so that $\mathcal{F}$ identifies P^{tar} as its unique minimizer. We choose the Lipschitz-regularized KL divergence over standard choices such as the KL divergence or the Wasserstein-1 distance since, unlike those, both the Lipschitz-regularized KL divergence and its variational derivative exist finitely at the population level and at the finite-sample level, with no additional assumptions on the target (Gu et al., 2024; Chen et al., 2025). The smoothed CVaR and its variational derivative likewise exist without any density assumptions. Thus, both $\mathcal{F}$ and its variational derivative are well defined for empirical distributions supported on finite samples, which allows us to *initialize the fine-tuning flow using samples from* P^{pre}, satisfying property **(I)**.

To place additional weight on the tail region, we use the smoothed CVaR at level α of the radial function $g(x) = \|x\|$, which is the average radius over the most extreme $(1-\alpha)$-fraction of the samples, and thus by definition focuses on the tail region. Moreover, the CVaR penalty depends on the target only through the single scalar statistic $\mathrm{CVaR}^{P^{\mathrm{tar}},g}_{\alpha,h}$, whose finite-sample estimation error will be shown to be independent of the dimension d of the target distribution, so the number

of extreme target samples the CVaR penalty requires does not grow with d. We derive the velocity field of the fine-tuning flow and show that the CVaR penalty contributes a velocity component that *focuses on the tail region*, mitigating its under-sampling, even for high-dimensional heavy-tailed targets, satisfying property **(II)**. Furthermore, the resulting velocity field of the fine-tuning flow is target-adaptive, thus satisfying property **(III)**. In particular, the velocity component from the CVaR penalization is determined by the CVaR discrepancy $\Delta C_h(Q; P^{\mathrm{tar}})$, and remains active whenever the discrepancy is non-zero. Thus, the magnitude of the velocity field, as well as the terminal time of the fine-tuning flow, is implicitly determined by the target.

We evaluate CVaR-GPA by fine-tuning seven pre-trained generative models: Lip-KL-GPA (Chen et al., 2025), Tail Transform Flow (TTF) (Hickling & Prangle, 2025), Tail-GAN (Cont et al., 2025), t-EDM (Pandey et al., 2025), Score-based Heavy-tailed Diffusion (SHD) (Liu et al., 2026), Denoising Lévy Probabilistic Models (DLPM) (Shariatian et al., 2025), and marginal Tail-Adaptive Flows (mTAF) (Laszkiewicz et al., 2022). Together they span normalizing flows, GANs, diffusion and score-based approaches as well as Wasserstein gradient flows. Figure 1 shows the global and tail errors of the seven pre-trained models, before and after fine-tuning using CVaR-GPA, on the daily streamflow in the Ohio River basin ($d = 64$) (U.S. Geological Survey, 2026), an extremely heavy-tailed target whose Hill tail indices (Hill, 1975) range from 1.05 to 3.34, with 51 of the 64 marginals in the infinite-variance regime. Both errors are reduced for every model after fine-tuning, by geometric-mean factors of $9.5\times$ globally and $5.4\times$ in the tail. Across four targets and seven pre-trained models, CVaR-GPA reduces errors in all but one of the 56 dataset–model–error metric combinations (Table 1), by geometric-mean factors of $14.0\times$ globally and $9.8\times$ in the tail, with a single set of hyperparameters and without any model-specific or target-specific tuning. These results empirically demonstrate how CVaR-GPA improves the accuracy of the pre-trained models and makes the fine-tuning robust across various models and datasets.

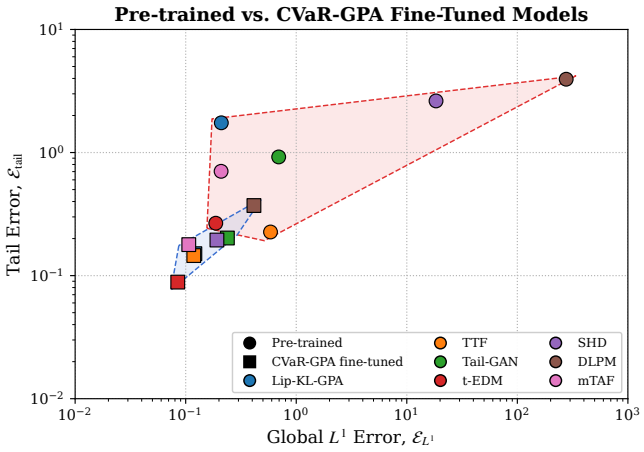

Figure 1: Global L^1 and tail errors of seven pre-trained models before (circles) and after (squares) fine-tuning via CVaR-GPA, on the daily streamflow in the Ohio River basin ($d = 64$) (U.S. Geological Survey, 2026) on a logarithmic axis. The models and the error metrics are given in Section 4.

The rest of the paper is organized as follows. Section 2 introduces the CVaR-penalized loss functional and discusses how the induced Wasserstein gradient flow and its velocity field realize **(I)**–**(III)**. Section 3 develops the corresponding particle algorithm CVaR-GPA, and Section 4 evaluates CVaR-GPA across four heavy-tailed targets (synthetic and real-world) and seven pre-trained generative models. Section 5 concludes the paper, while additional related work is discussed in the Appendix.

2 Target-adaptive fine-tuning flow

The loss functional $\mathcal{F}$ in Equation 1 comprises two main components: the Lipschitz-regularized KL divergence and the smoothed CVaR.

Lipschitz-regularized KL divergence Introduced in Dupuis & Mao (2022) and subsequently developed in Birrell et al. (2022) as a computational tool for generative modeling, the Lipschitz-regularized KL divergence is defined via the infimal convolution:

$$D^L_{\mathrm{KL}}(Q\|P^{\mathrm{tar}}) = \inf_{\gamma\in\mathcal{P}(\mathbb{R}^d)} \left\{D_{KL}(\gamma\|P^{\mathrm{tar}}) + L\cdot\mathcal{W}_1(Q,\gamma)\right\}, \tag{2}$$

where D_{KL} and $\mathcal{W}_1$ denote the KL divergence and the Wasserstein-1 metric, respectively. By definition, the minimizer γ^* of Equation 2 can be viewed as a reweighting of the target measure P^{tar} through the KL divergence, while the Wasserstein-1 metric measures the cost of transporting mass from the intermediate measure γ^* to Q (see Birrell et al. (2022, Section 3) for a detailed discussion of the mass redistribution/transport interpretation). Furthermore, from Dupuis & Mao (2022), the Lipschitz-regularized KL divergence defined by Equation 2 has a variational representation:

$$D^L_{\mathrm{KL}}(Q\|P^{\mathrm{tar}}) = \sup_{\phi\in\Gamma_L}\{\mathbb{E}_Q[\phi] - \log(\mathbb{E}_{P^{\mathrm{tar}}}[e^{\phi}])\}, \tag{3}$$

where Γ_L is the space of L-Lipschitz functions. Let $\mathcal{S} := \mathrm{supp}(Q)\cup\mathrm{supp}(P^{\mathrm{tar}})$, where $\mathrm{supp}(Q)$ and $\mathrm{supp}(P^{\mathrm{tar}})$ denote the supports of Q and P^{tar}, respectively. From Chen et al. (2025, Theorem 3), the optimizer of Equation 3 exists and is unique (up to a constant) on $\mathcal{S}$; we denote it by $\phi^*_{\mathcal{S}}: \mathcal{S}\to\mathbb{R}$. Extending $\phi^*_{\mathcal{S}}$ to $\mathbb{R}^d$ via $\phi^*(x) := \inf_{y\in\mathcal{S}}\left\{\phi^*_{\mathcal{S}}(y) + L\|x-y\|\right\}$ for $x\in\mathbb{R}^d$, we obtain $\phi^*\in\Gamma_L$ with $\phi^*|_{\mathcal{S}} = \phi^*_{\mathcal{S}}$, and for all $x\in\mathbb{R}^d$,

$$\frac{\delta D^L_{\mathrm{KL}}(Q\|P^{\mathrm{tar}})}{\delta Q}(x) = \phi^*(x).$$

For empirical measures supported on finite samples, both $D^L_{\mathrm{KL}}(Q\|P^{\mathrm{tar}})$ and its first variational derivative exist finitely (see Chen et al. (2025, Theorem 2, Theorem 3)). Due to the scarcity of samples in the tail region, the velocity field induced by the Lipschitz-regularized KL divergence alone, often vanishes before the tail region is adequately learned. Thus, we penalize the Lipschitz-regularized KL divergence via CVaR to place additional weight on the tail region.

Conditional Value-at-Risk Let $g:\mathbb{R}^d\to[0,\infty)$ be a risk function. For $Q\in\mathcal{Q}^g := \{Q\in\mathcal{P}(\mathbb{R}^d) : \mathbb{E}_Q[g] < \infty\}$, the CVaR of g at level $\alpha\in(0,1)$ is the average of g over the extreme $(1-\alpha)$-fraction of outcomes: $\mathrm{CVaR}^{Q,g}_\alpha = \mathbb{E}_Q[g\,|\,g\geq q^{Q,g}_\alpha]$ when the law $g_\#Q$ of $g(X)$, $X\sim Q$, is continuous, with $q^{Q,g}_\alpha$ the α-quantile of $g_\#Q$. Therefore, CVaR, by definition, focuses on the tail region of a given distribution, and hence is a natural choice for learning heavy-tailed targets. The Rockafellar–Uryasev formulation (Rockafellar & Uryasev, 2002) is a variational formulation of CVaR, which defines for every $Q\in\mathcal{Q}^g$:

$$\mathrm{CVaR}^{Q,g}_\alpha = \inf_{y\in\mathbb{R}} F^{Q,g}_\alpha(y) := \inf_{y\in\mathbb{R}}\left\{y + \frac{1}{1-\alpha}\mathbb{E}_Q[(g(\cdot)-y)^+]\right\}, \tag{4}$$

where $u^+ = \max\{0,u\}$ for $u\in\mathbb{R}$. As shown in Theorem D.5 in Appendix D, CVaR, in general, does not have a variational derivative, which is required to define Wasserstein gradient flows. Thus, we instead consider the smoothed CVaR introduced in Kouri & Surowiec (2016). We refer to Appendix D for detailed discussions on penalization via (unsmoothed) CVaR, using Clarke's generalized subgradients, and how it compares to the smoothed CVaR penalization. For a smoothing parameter $h>0$, the smoothed CVaR is defined by

$$\mathrm{CVaR}^{Q,g}_{\alpha,h} = \inf_{y\in\mathbb{R}} F^{Q,g}_{\alpha,h}(y) := \inf_{y\in\mathbb{R}}\left\{y + \frac{1}{1-\alpha}\,\mathbb{E}_Q\left[r_h(g-y)\right]\right\}, \tag{5}$$

where $r_h(u) := u^+ + h\log\left(1+e^{-|u|/h}\right)$ is the softplus smoothing of u^+. We denote by $q^{Q,g}_{\alpha,h}$ the minimizer in Equation 5, unique since $F^{Q,g}_{\alpha,h}(\cdot)$ is strictly convex. By Proposition B.4, the variational derivative of the smoothed CVaR exists for any $Q\in\mathcal{Q}^g$, including empirical measures supported on finite samples, and is given by

$$\frac{\delta}{\delta Q}\mathrm{CVaR}^{Q,g}_{\alpha,h} = \frac{1}{1-\alpha}r_h\left(g - q^{Q,g}_{\alpha,h}\right).$$

Throughout this work, CVaR-penalization refers to penalization via the smoothed CVaR.

Defining CVaR penalization through a scalar risk function g is a crucial design feature: the CVaR penalization depends only on the target distribution via the one-dimensional pushforward measure $g_{\#}P^{\mathrm{tar}}$, through the scalar tail statistic $\mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}$, whose estimation error is independent of d and h, as shown in the following lemma.

Lemma 2.1. *Let $\widehat{P^{\mathrm{tar}}}$ be the empirical measure of N i.i.d. samples of P^{tar}, and assume that $g_{\#}P^{\mathrm{tar}}$ admits a finite q^{th} moment for some $q > 1$. Then the finite-sample estimation error $\mathbb{E}_{(P^{\mathrm{tar}})^{\otimes N}}|\mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}|$ is of order $O(N^{-\min\{\frac{1}{2},\frac{q-1}{q}\}})$ (for $q = 2$, with q replaced by any $q' \in (1,2)$), with an implicit constant independent of the dimension d and the smoothing parameter $h > 0$.*

The precise statement and the proof of Lemma 2.1 are provided in Appendix B.4. Consequently, the number of samples needed to estimate the target CVaR to a prescribed accuracy does not grow with d, making fine-tuning via CVaR-penalization scalable to high dimensions. CVaR penalization in the loss functional $\mathcal{F}$ thus places an additional weight on the tail region, at an estimation cost that does not grow with d.

On the other hand, the Lipschitz-regularized KL divergence $D_{\mathrm{KL}}^{L}(Q\|P^{\mathrm{tar}})$ is equally essential to the CVaR-penalized loss functional. For every $Q \in \mathcal{Q}^g$ with a finite first moment, $\mathcal{F}(Q;P^{\mathrm{tar}})$ satisfies the divergence property, i.e., $\mathcal{F}(Q;P^{\mathrm{tar}}) \geq 0$, and the equality holds if and only if $Q = P^{\mathrm{tar}}$ (see Proposition B.3 in the Appendix), making P^{tar} the unique global minimizer of $\mathcal{F}(\cdot\,;P^{\mathrm{tar}})$. The term $(\Delta C_h(Q;P^{\mathrm{tar}}))^2$ alone does not satisfy this property, since infinitely many distributions Q share the same smoothed CVaR as P^{tar}. It is $D_{\mathrm{KL}}^{L}(Q\|P^{\mathrm{tar}})$ that endows $\mathcal{F}$ with the divergence property, and thereby ensures that its minimizer uniquely identifies the target distribution. Moreover, since it compares Q and P^{tar} directly as a divergence, $D_{\mathrm{KL}}^{L}(Q\|P^{\mathrm{tar}})$ also captures structural properties of P^{tar} that the scalar statistic $\mathrm{CVaR}_{\alpha,h}^{Q,g}$ cannot reflect. Thus, the two components of the loss functional play distinct and complementary roles.

We have the following result, whose proof can be found in the Appendix.

Theorem 2.2. *Let $g \in C^1$ and $Q \in \mathcal{Q}^g$, and $h > 0$. The Wasserstein gradient flow of $\mathcal{F}(Q;P^{\mathrm{tar}})$ is governed by the transport PDE:*

$$\partial_t Q_t + \nabla\cdot(Q_t v_t(x)) = 0, \quad v_t(x) = -\nabla_x\Big(\frac{\delta\mathcal{F}(Q;P^{\mathrm{tar}})}{\delta Q}\Big|_{Q=Q_t}\Big), \quad t>0, \quad Q_0 = P^{\mathrm{pre}}, \tag{6}$$

where Q_t denotes the evolving distribution at time t and the variational derivative of $\mathcal{F}$ is given by

$$\frac{\delta\mathcal{F}(Q;P^{\mathrm{tar}})}{\delta Q} = \phi^* - \frac{2\lambda}{1-\alpha}\,\Delta C_h(Q;P^{\mathrm{tar}})\,r_h\big(g - q_{\alpha,h}^{Q,g}\big). \tag{7}$$

The smoothed CVaR-penalized Wasserstein gradient flow is thus governed by Equation 6, with the velocity field

$$v_{Q,h}(x) = \underbrace{-\nabla_x\phi^*(x)}_{\text{from } D_{\mathrm{KL}}^{L}(Q\|P^{\mathrm{tar}})} + \underbrace{\lambda\,\frac{2}{1-\alpha}\,\Delta C_h(Q;P^{\mathrm{tar}})\,\sigma\Big(\frac{g(x)-q_{\alpha,h}^{Q,g}}{h}\Big)\nabla_x g(x)}_{\hat{v}_{Q,h}(x)\text{: velocity component from CVaR-penalization}}, \tag{8}$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$. The logistic factor $\sigma\big(\frac{g(x)-q_{\alpha,h}^{Q,g}}{h}\big)$ is a smooth surrogate for the indicator function of $\{g(x) > q_{\alpha,h}^{Q,g}\}$, and thus localizes the velocity component $\hat{v}_{Q,h}$ to the tail region. In what follows, we fix the risk function as $g(x) = \|x\|$, which captures extremes in every direction through a scalar, and hence is a natural choice for g. The resulting velocity field is given by Equation 8 with $\nabla_x g(x) = \frac{x}{\|x\|}$, adopting the convention $\frac{x}{\|x\|} := 0$ and $v_{Q,h}(x) = -\nabla_x\phi^*(x)$ at $x = 0$.

In CVaR-GPA, the velocity component from the Lipschitz-regularized KL divergence is approximated via neural network approximations, similar to Chen et al. (2025); Gu et al. (2024), and learns the joint structure of the target P^{tar} due to its divergence property, which the scalar CVaR discrepancy alone cannot capture. For every $Q \in \mathcal{Q}^g$ with a finite first moment, including empirical measures supported on finite samples, $\mathcal{F}(Q;P^{\mathrm{tar}})$ and its variational derivative (Equation 7) are well defined, allowing us to initialize the fine-tuning flow (Equation 6) with samples of the pre-trained

model P^{pre}, satisfying property **(I)**. The number of samples required to estimate the velocity component $\hat{v}_{Q,h}$ does not scale with d, since $\hat{v}_{Q,h}$ depends on P^{tar} only through the scalar $\mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}$, which, as established in Lemma 2.1, can be estimated with a finite-sample error bound independent of the target dimension d. Since $\sigma(t) > 0$, $\hat{v}_{Q,h}(x) \neq 0$ whenever $\Delta C_h(Q; P^{\mathrm{tar}}) \neq 0$ and $x \neq 0$, irrespective of how many particles lie in the tail region: the CVaR component remains active in the under-sampled tail. Thus, the CVaR-penalization contributes a *dimension-free* velocity component $\hat{v}_{Q,h}$ that *remains active in the tail region*, satisfying property **(II)**.

The velocity field $v_{Q,h}$ is *target-adaptive* via the CVaR discrepancy $\Delta C_h(Q; P^{\mathrm{tar}})$, satisfying property **(III)**. The magnitude of $\hat{v}_{Q,h}$ is proportional to $|\Delta C_h(Q; P^{\mathrm{tar}})|$. That is, the CVaR penalization supplies a self-regulating velocity field, driven by the CVaR discrepancy. We further note that the velocity component $-\nabla_x \phi^*(\cdot)$ is merely bounded but not necessarily Lipschitz continuous, while the velocity component $\hat{v}_{Q,h}$ is locally Lipschitz on $\mathbb{R}^d \setminus \{0\}$ but admits no global Lipschitz bound. Thus, the velocity field $v_{Q,h}$ does not adhere to the global Lipschitz constraint of Jaini et al. (2020).

3 CVaR-PENALIZED GENERATIVE PARTICLE ALGORITHM

The smoothed CVaR-penalized Wasserstein gradient flow describes the time-evolving distribution Q_t of a particle system. Equivalently, the dynamics of the particle system can be described by the Lagrangian formulation of the transport/variational PDE Equation 6 given by

$$\frac{d}{dt} Y_t = v_{Q_t,h}(Y_t), \qquad Y_0 \sim P^{\mathrm{pre}}, \qquad \text{where } Y_t \sim Q_t, \tag{9}$$

with the velocity field $v_{Q_t,h}$ given by Equation 8. This motivates a neural particle algorithm, analogous to the Generative Particle Algorithm (GPA) developed in Gu et al. (2024), but with an additional CVaR penalty. We build the CVaR-penalized Generative Particle Algorithm (CVaR-GPA) via an Euler discretization of Equation 9, by replacing the probability measures with their empirical counterparts.

The update scheme of CVaR-GPA We denote by $\widehat{Q}_k$ the empirical measure of the generated distribution $\{Y_k^{(i)}\}_{i=1}^M$ at the kth iteration, and by $\widehat{P^{\mathrm{tar}}}$ the empirical measure of the target samples $\{X^{(j)}\}_{j=1}^N$. Let Γ_L^{NN} be a neural network approximation, with spectral normalization (Miyato et al., 2018), of the function space Γ_L of L-Lipschitz functions. Because $\widehat{Q}_k$ evolves over iterations, at each step the function ϕ_k^* is re-obtained by solving the variational problem Equation 3 over the approximated function space Γ_L^{NN}. The velocity contribution $-\nabla_x \phi_k^*(Y_k^{(i)})$ from the Lipschitz-regularized KL divergence is evaluated by automatic differentiation of the network at the particle positions. Thus, for a chosen step size Δt, the kth iteration of the update scheme is

$$\begin{aligned}
&Y_{k+1}^{(i)} = Y_k^{(i)} + \Delta t\, v_k^{(i)}, \quad v_k^{(i)} := -\nabla_x \phi_k^*(Y_k^{(i)}) + \lambda\, b_k^{(i)}, \quad Y_0^{(i)} \sim P^{\mathrm{pre}}, \quad i = 1, \ldots, M, \\
&\phi_k^* = \arg\max_{\phi \in \Gamma_L^{\mathrm{NN}}} \Big\{ \frac{1}{M} \sum_{i=1}^M \phi(Y_k^{(i)}) - \log\Big(\frac{1}{N} \sum_{j=1}^N e^{\phi(X^{(j)})} \Big) \Big\}, \\
&b_k^{(i)} = \frac{2\,\Delta_k}{1-\alpha}\, w_k^{(i)}\, \frac{Y_k^{(i)}}{\|Y_k^{(i)}\|}, \quad \Delta_k := \mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{\widehat{Q}_k,g}, \quad w_k^{(i)} := \sigma\Big(\frac{\|Y_k^{(i)}\| - q_{\alpha,h}^{\widehat{Q}_k,g}}{h} \Big).
\end{aligned} \tag{10}$$

The necessary tail statistics for the kth update, namely $q_{\alpha,h}^{\widehat{Q}_k,g}$, the weights $w_k^{(i)}$, and the empirical CVaR discrepancy Δ_k, are computed via Algorithm 1 in Appendix B. In contrast to ϕ_k^*, whose approximation uses a neural network, these empirical tail statistics are computed directly from the particle samples and require no neural network approximation. The fine-tuning flow terminates when its velocity field dies out, i.e., when $v_{Q_{T^*},h} = 0$ for some terminal time $T^* > 0$. Hence, one can use the empirical estimate of the velocity field, or the empirical estimate of the total kinetic energy of the generative particles at the kth iteration, given by $\mathcal{K}_k := \frac{1}{2M} \sum_{i=1}^M \|v_k^{(i)}\|_2^2$, to define a stopping criterion for CVaR-GPA. We terminate the algorithm when $\mathcal{K}_k < \epsilon$ for a given threshold ϵ.

The update scheme Equation 10, together with the stopping criterion, describes CVaR-GPA, which is stated in full as Algorithm 2 in Appendix B. CVaR-GPA inherits the properties **(I)**–**(III)** of the underlying flow at the particle level. The update scheme (Equation 10) is initialized with samples from

P^{pre}, without requiring access to the architecture of pre-trained models. The CVaR velocity component $b_k^{(i)}$ places additional weight on the tail: since $w_k^{(i)} > 0$, every tail particle receives a nonzero displacement whenever $\Delta_k \neq 0$, however few particles lie in the tail. The CVaR velocity component $b_k^{(i)}$ adapts to the target via the empirical CVaR discrepancy Δ_k, and the stopping criterion results in an algorithm whose architecture depth is implicitly determined by P^{tar}. Thus, CVaR-GPA does not require the tail index or an estimate of the tail index as an input, making CVaR-GPA a tail-agnostic algorithm. Together, these properties improve the accuracy in learning heavy-tailed targets and make CVaR-GPA robust across various pre-trained models and datasets, as we demonstrate in Section 4.

4 Numerical Experiments

We evaluate CVaR-GPA on both synthetic and real-world datasets that assess its robustness across varying tail indices and targets with correlated marginals, and its scalability to high dimensions. The experiments demonstrate how the properties **(I)–(III)** embedded in the fine-tuning flow improve the accuracy of the pre-trained models and make the fine-tuning robust across pre-trained models of different architectures and datasets with heterogeneous tail behaviors.

Datasets We first assess the tail-agnostic performance of CVaR-GPA on a *2-dimensional isotropic Student-t* distribution with tail indices $\nu \in \{1, 1.2, 1.5, 1.8\}$, where a smaller ν corresponds to a heavier power-law tail. CVaR-GPA is applied to all four targets using identical hyperparameters, without prior knowledge of ν. We next consider *Neal's funnel distribution* (Neal, 2003), a standard benchmark for heavy-tailed generative models (Jaini et al., 2020; Pandey et al., 2025), characterized by strongly heterogeneous and dependent marginal behavior. Both synthetic experiments use $N = 5000$ target samples. We then evaluate CVaR-GPA on two real-world, high-dimensional datasets with heterogeneous heavy tails, drawn from domains in which accurate modeling of extremes is consequential. The first is the *Fama–French 25 monthly portfolios* (Fama & French, 1993) ($d = 25$, $N = 1182$ months), a canonical benchmark in empirical asset pricing, with marginal Hill tail-index estimates (Hill, 1975) ranging from 2.2 to 3.2. The second is *daily streamflow* (U.S. Geological Survey, 2026) from the 64 gauges with the smallest drainage areas among the 312 Ohio River basin gauges having complete records over 2005–2024 ($d = 64$, $N = 7305$ days). Streamflow extremes directly inform flood-risk assessment and hydraulic design, making accurate modeling of tail behavior particularly consequential. Its estimated marginal tail indices range from 1.05 to 3.34, with 51 of the 64 gauges having tail indices below 2, and therefore lying in the infinite-variance regime. Streamflow thus provides a particularly demanding high-dimensional real-world benchmark, with extremely heavy-tailed marginals where the tail is especially difficult to learn from finite samples.

Pre-trained models For each dataset, we fine-tune seven pre-trained generative models: Lip-KL-GPA (Chen et al., 2025), TTF (Hickling & Prangle, 2025), Tail-GAN (Cont et al., 2025), t-EDM (Pandey et al., 2025), SHD (Liu et al., 2026), DLPM (Shariatian et al., 2025), and mTAF (Laszkiewicz et al., 2022). All models are trained on the same target samples using the settings reported by their respective authors. We note here that Tail-GAN is designed mainly to reproduce tail risk statistics of financial portfolios with heavy-tailed distributions rather than to learn the underlying distribution. We nonetheless include it because its training objective also incorporates CVaR.

Parameter selection The number of generated particles is set to be equal to the number of target samples, i.e., $M = N$. The same hyperparameters $(\lambda, \alpha, h, L, \Delta t) = (4(1-\alpha) \times 10^{-3}, 1 - \frac{3}{N}, 0.5, 0.125, 25)$ are used across all four datasets and all seven pre-trained models. The order of λ is chosen at the level of $1 - \alpha$, so that the coefficient of the CVaR-induced velocity component in Equation 8 is independent of N. The choice of α is discussed in detail in Appendix B.1. No hyperparameter is adjusted to the target distribution or the model being fine-tuned; in particular, CVaR-GPA does not use estimates of the target tail-decay rate. CVaR-GPA is terminated when the empirical kinetic energy falls below the prescribed threshold $\epsilon = 10^{-10}$, as described in Algorithm 2, with a maximum of 20,000 iterations imposed as a computational safeguard.

Error metrics Standard error metrics weight all regions of the domain equally and can therefore understate tail error, since the extreme tail carries only a small fraction of the total probability mass. A model may therefore achieve low global error while still failing to capture the tail accurately. Thus,

Table 1: Error-reduction factors (pre-trained error divided by fine-tuned error) for seven models after CVaR-GPA fine-tuning on the four targets: the 2-dimensional Cauchy distribution, Neal's funnel, Fama–French 25, and streamflow. The only increase in error (1 out of 56 cases) is shown in gray. The last row reports the geometric mean across the seven models for each target. The corresponding error values across the seven pre-trained models are reported in Appendix C.1.

pre-trained	Cauchy		Neal		Fama–French 25		Streamflow	
model	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$
Lip-KL-GPA	6.6×	67.3×	4.5×	4.6×	4.2×	36.1×	1.7×	11.6×
TTF	2.5×	0.9×	13.6×	16.0×	32.7×	6.2×	4.9×	1.6×
Tail-GAN	33.0×	130.6×	33.6×	77.5×	9.6×	20.1×	2.9×	4.6×
t-EDM	5.1×	5.7×	7.8×	6.2×	3.5×	4.1×	2.2×	3.0×
SHD	5.4×	2.7×	73.9×	19.0×	34.0×	20.7×	95.9×	13.5×
DLPM	35.0×	1.6×	267.4×	10.5×	468.2×	25.0×	667.7×	10.6×
mTAF	4.8×	5.2×	24.8×	20.7×	11.9×	28.8×	2.0×	3.9×
geom. mean	**8.2**×	**7.2**×	**25.9**×	**14.7**×	**18.9**×	**16.1**×	**9.5**×	**5.4**×

we use both standard and tail-sensitive metrics for evaluation. The **global L^1 error**, $\mathcal{E}_{L^1}$, is the L^1 distance between the complementary CDFs (CCDFs) of the risk values under the generated and target empirical distributions, $\widehat{Q}$ and $\widehat{P^{\text{tar}}}$. The **tail error**, $\mathcal{E}_{\text{tail}}$, is the mean log-CCDF discrepancy between $\widehat{Q}$ and $\widehat{P^{\text{tar}}}$ over the tail region beyond the 0.95 quantile of $\widehat{P^{\text{tar}}}$. Both metrics are defined precisely in Appendix B.2.

Error reduction after fine-tuning Table 1 reports the global and tail error-reduction factors across the seven pre-trained models for four representative targets: the Cauchy distribution ($\nu = 1$) from the Student-t family, Neal's funnel, Fama–French 25, and daily streamflow. Because the error-reduction factors are multiplicative ratios and span several orders of magnitude, we summarize them across models using the geometric mean. The corresponding error values are reported in Appendix C.1. Across these 56 dataset–model–error metric combinations, the global and tail geometric mean reductions are $14.0\times$ and $9.8\times$, respectively. Crucially, across these 56 instances, CVaR-GPA reduces errors in all but one case, without any model-specific modifications, highlighting the *architecture-agnostic* nature of the GPA framework. The one case where the error increases occurs for the tail error of TTF on the Cauchy target, where the pre-trained model already achieves a relatively small tail error of 0.138, leaving little room for further improvement; consequently, fine-tuning provides no additional benefit and instead leads to a marginal increase of 0.01.

The empirical results validate features that CVaR-GPA inherits from its underlying mathematical framework. Table 2 reports the global and tail error-reduction factors after CVaR-GPA fine-tuning of the seven pre-trained models on the **2-dimensional Student-t family**, across four tail indices, $\nu \in \{1, 1.2, 1.5, 1.8\}$. Crucially, these error reductions are obtained with the same hyperparameters across the entire family: CVaR-GPA requires no tail-index-specific adaptation, since the velocity field Equation 8, driving CVaR-GPA, implicitly adapts to the target. Figure 2 presents samples from two pre-trained models, Tail-GAN and mTAF, for which we obtain the largest tail-error reduction factors, for **Neal's funnel distribution**, before and after fine-tuning. In each case, CVaR-GPA recovers the characteristic funnel geometry, a key feature underlying Neal's funnel's use as a challenging benchmark, from pre-trained samples that fail to accurately reproduce this structure. CVaR-GPA operates directly on the joint target distribution, through the Lipschitz-regularized KL divergence component of its loss functional, as opposed to learning each marginal separately, and is thus expected to perform well on anisotropic targets with correlated marginal distributions, as

Table 2: Geometric mean, over the seven pre-trained models, of the error-reduction factors after CVaR-GPA fine-tuning on the 2-dimensional isotropic Student-t family, for each tail index ν.

tail index	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$
$\nu = 1$ (Cauchy)	8.2×	7.2×
$\nu = 1.2$	20.6×	11.2×
$\nu = 1.5$	24.1×	13.4×
$\nu = 1.8$	18.5×	10.2×
geom. mean	**16.6**×	**10.3**×

demonstrated empirically by Figure 2. Corresponding results for other pre-trained models, together with per-dimension error reductions, are in Appendix C.3.

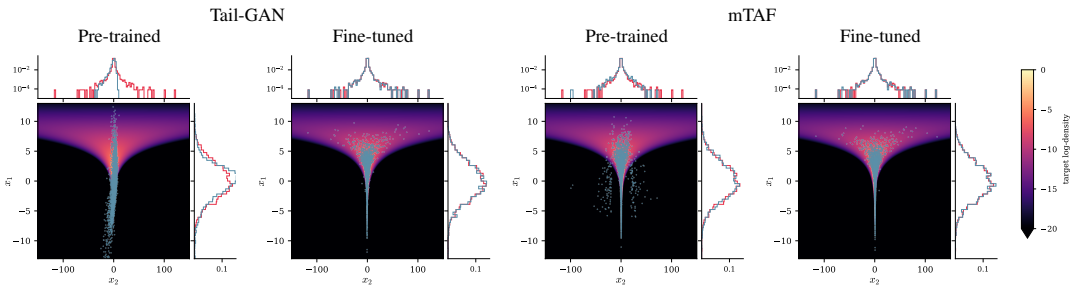

Figure 2: Neal's funnel distribution: pre-trained and CVaR-GPA fine-tuned samples, given by blue particles, for the two models with the largest tail-error reduction factors, Tail-GAN and mTAF. The target log-density is shown in the background with the shared color scale at right. Joint error-reduction factors $(\mathcal{E}_{\text{tail}}, \mathcal{E}_{L^1})$ are $(77.5\times,\ 33.6\times)$ for Tail-GAN and $(20.7\times,\ 24.8\times)$ for mTAF (best viewed when zoomed in; see Appendix C.3 for details).

On the two **high-dimensional real-world targets**, fine-tuning improves both global and tail error of every pre-trained model (Table 1). On Fama–French 25, the geometric-mean reductions are $18.9\times$ in global error and $16.1\times$ in tail error; on streamflow, they are $9.5\times$ and $5.4\times$, respectively. These datasets span markedly different heavy-tail regimes: the Fama–French marginals have estimated tail indices between 2.2 and 3.2 (Hill, 1975), whereas streamflow extends down to 1.05, with 51 of 64 marginals in the infinite-variance regime. Thus, the improvements persist from moderately heavy-tailed to extremely heavy-tailed targets. Per-marginal results for Fama–French 25 and streamflow are presented in Appendices C.5 and C.6. The tail statistic driving the CVaR velocity component is a scalar, so by Lemma 2.1 its estimation does not degrade with d; this is what allows CVaR-GPA to be applied unchanged at $d = 25$ and $d = 64$, making the method scalable to high dimensions.

5 Conclusion

We develop CVaR-GPA, a tail-agnostic generative particle algorithm for fine-tuning pre-trained models toward heavy-tailed targets. It is built on the Wasserstein gradient flow of the Lipschitz-regularized KL divergence penalized by a smoothed Conditional Value-at-Risk discrepancy. The loss functional is designed so that the resulting flow satisfies properties **(I)**–**(III)**. With hyperparameters fixed across all experiments and not tuned to particular targets or pre-trained models, CVaR-GPA improves the accuracy of pre-trained models on heavy-tailed targets with heterogeneous tail behaviors in 55 out of 56 dataset–model–error metric combinations. The global and tail geometric-mean error reductions are $14.0\times$ and $9.8\times$, respectively. Each of the three properties **(I)**–**(III)** is reflected in these results. Since the flow is initialized directly from pre-trained samples **(I)**, CVaR-GPA is architecture-agnostic, which explains its robustness across all seven pre-trained models. The CVaR penalty keeps the velocity field active in the tail, even for high-dimensional target distributions **(II)**, allowing CVaR-GPA to scale unchanged to the high-dimensional real-world targets with $d = 25$ and $d = 64$. The target-adaptive velocity field (Equation 8) implicitly determines, for each target, the depth of the transport map required to learn it **(III)**. It thereby removes the need for tail-index-specific tuning across the four datasets with heterogeneous tail behaviors.

While the radial risk function and the smoothed CVaR are natural choices for learning heavy-tailed targets, the same design principle, embedding desired properties into the flow through the choice of the loss functional, applies to other couplings of (possibly anisotropic) risk functions g with risk measures (in particular spectral risk measures (Acerbi & Simonetti, 2002), whose weight functions can be tailored to achieve task-specific goals) for specific generative tasks, such as learning multi-modal target distributions. Moreover, the cost of re-running the particle flow for new targets can be amortized by adapting to the CVaR-penalized flow the force-matching strategies developed in Cheng et al. (2025) to learn continuous gene expression dynamics from GPA trajectories.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Assisting with translation does not apply to this work. Additionally, we used generative AI tools for drafting parts of the paper, identifying relevant literature, and editing the paper to improve readability. We reviewed all AI-assisted content and checked it for correctness, and we take full responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors believe that this paper submission does not raise questions regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

The source code for the algorithm can be found at `https://osf.io/deph9/overview?view_only=0aa7262aef4b4d9ba673957820e97eed`. The hyperparameters used and other experimental details are stated in Section 4. For all the theoretical results, the assumptions and complete proofs of the claims are included in the appendix. References for all the datasets used are provided.

REFERENCES

Carlo Acerbi and Prospero Simonetti. Portfolio optimization with spectral measures of risk. *arXiv preprint cond-mat/0203607*, 2002.

Carlo Acerbi, Claudio Nordio, and Carlo Sirtori. Expected shortfall as a tool for financial risk management. *arXiv preprint cond-mat/0102304*, 2001.

Hansjörg Albrecher and Søren Asmussen. Ruin probabilities and aggregrate claims distributions for shot noise Cox processes. *Scandinavian Actuarial Journal*, 2006(2):86–110, 2006.

Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. EV-GAN: Simulation of extreme events with ReLU neural networks. *Journal of Machine Learning Research*, 23(150):1–39, 2022.

Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. ExceedGAN: simulation above extreme thresholds using generative adversarial networks. *Extremes*, pp. 1–23, 2026.

Siddharth Bhatia, Arjit Jain, and Bryan Hooi. ExGAN: Adversarial generation of extreme samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6750–6758, 2021.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f,Gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022.

Kai Chang and Themistoklis P Sapsis. Extreme event aware (η-) learning. *Nature Communications*, 2026.

Sapana Chaudhary, Ujwal Dinesha, Dileep Kalathil, and Srinivas Shakkottai. Risk-averse fine-tuning of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Ziyu Chen, Hyemin Gu, Markos A Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Robust generative learning with Lipschitz-regularized α-divergences allows minimal assumptions on target distributions. *Information and Inference: A Journal of the IMA*, 14(4):iaaf028, 2025.

Yu-Chen Cheng et al. PROFET predicts continuous gene expression dynamics from scRNA-seq data to elucidate heterogeneity of cancer treatment responses. *bioRxiv*, 2025. doi: 10.1101/2025.06.27.662030.

Pasquale Cirillo and Nassim Nicholas Taleb. Tail risk of contagious diseases. *Nature Physics*, 16 (6):606–613, 2020.

Frank H Clarke. *Optimization and nonsmooth analysis*. Society for industrial and Applied Mathematics, 1990.

Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-GAN: Learning to simulate tail risk scenarios. *Management Science*, 2025.

Paul Dupuis and Yixiang Mao. Formulation and properties of a divergence used to compare probability measures without absolute continuity. *ESAIM: Control, Optimisation and Calculus of Variations*, 28:10, 2022.

Paul Embrechts and Noël Veraverbeke. Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics*, 1(1):55–72, 1982.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*, volume 33 of *Applications of Mathematics*. Springer, Berlin, 1997.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. doi: 10.1016/0304-405X(93)90023-5.

Mohammad Farazmand and Themistoklis P Sapsis. Closed-loop adaptive control of extreme events in a turbulent flow. *Physical Review E*, 100(3):033110, 2019.

Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2 edition, 2013. doi: 10.1007/978-1-4614-7101-1.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Patricia Grossi and Howard Kunreuther. *Catastrophe modeling: a new approach to managing risk*, volume 25. Springer Science & Business Media, 2005.

Hyemin Gu, Panagiota Birmpa, Yannis Pantazis, Luc Rey-Bellet, and Markos A Katsoulakis. Lipschitz-regularized gradient flows and generative particle algorithms for high-dimensional scarce data. *SIAM Journal on Mathematics of Data Science*, 6(4):1205–1235, 2024.

Yunrui Guan, Krishna Balasubramanian, and Shiqian Ma. Mirror flow matching with heavy-tailed priors for generative modeling on convex domains. In *International Conference on Learning Representations*, 2026.

Tennessee Hickling and Dennis Prangle. Flexible tails for normalizing flows. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975. doi: 10.1214/aos/1176343247.

Todd Huster, Jeremy Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi O Leslie, Cho-Yu Jason Chiang, and Vyas Sekar. Pareto GAN: Extending the representational power of gans to heavy-tailed distributions. In *International Conference on Machine Learning*, pp. 4523–4532. PMLR, 2021.

Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz triangular flows. In *International Conference on Machine Learning*, pp. 4673–4681. PMLR, 2020.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Leonid Kantorovich and Gennady S. Rubinstein. On a space of totally additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

Masako Kishida. Risk-aware stability, ultimate boundedness, and positive invariance. *IEEE Transactions on Automatic Control*, 69(1):681–688, 2024.

Drew P Kouri and Thomas M Surowiec. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization*, 26(1):365–396, 2016.

Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Marginal tail-adaptive normalizing flows. In *International Conference on Machine Learning*, pp. 12020–12048. PMLR, 2022.

Haoyu Liu, Tingyu Zhu, Nanshan Jia, Jinghai He, and Zeyu Zheng. Learning to simulate from heavy-tailed distribution via diffusion model. *Operations Research*, 2026. doi: 10.1287/opre.2024.1450.

Christopher W Lynn, Caroline M Holmes, and Stephanie E Palmer. Heavy-tailed neuronal connectivity arises from Hebbian self-organization. *Nature Physics*, 20(3):484–491, 2024.

Bruno Merz, Stefano Basso, Svenja Fischer, David Lun, Günter Blöschl, Ralf Merz, Björn Guse, Alberto Viglione, Sergiy Vorogushyn, Elena Macdonald, et al. Understanding heavy tails of flood peak distributions. *Water Resources Research*, 58(6):e2021WR030506, 2022.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Jayakrishnan Nair, Adam Wierman, and Bert Zwart. *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*, volume 53 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2022.

Radford M Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001. doi: 10.1081/PDE-100002243.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *International Conference on Learning Representations*, 2025.

R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & Finance*, 26(7):1443–1471, 2002.

Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Flow density control: Generative optimization beyond entropy-regularized fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Dario Shariatian, Umut Simsekli, and Alain Oliviero Durmus. Denoising Lévy probabilistic models. In *International Conference on Learning Representations*, 2025.

U.S. Geological Survey. USGS Water Data for the Nation: U.S. Geological Survey National Water Information System database, 2026. Accessed September, 2026, at `https://doi.org/10.5066/F7P55KJN`.

Zifan Wang, Riccardo De Santi, Xiaoyu Mo, Michael M Zavlanos, Andreas Krause, and Karl H Johansson. Efficient tail-aware generative optimization via flow model fine-tuning. *arXiv preprint arXiv:2602.16796*, 2026.

APPENDIX

CONTENTS

A Related work **14**

B Supplementary material on the smoothed CVaR and CVaR-GPA **14**
B.1 Choice of the level α 14
B.2 Error metrics 16
B.3 Algorithms 16
B.4 Proofs of the results concerning the smoothed CVaR 16

C Numerical Experiments **22**
C.1 Full tables of errors corresponding to Table 1 22
C.2 Student-t family 23
C.3 Neal's funnel distribution 23
C.4 Per-marginal results 25
C.5 Fama–French 25: per-portfolio results 25
C.6 Streamflow: per-gauge results 26

D CVaR-penalized Wasserstein gradient flows with Clarke's generalized subgradients **31**
D.1 Clarke's generalized subgradients 31
D.2 Velocity field of the CVaR-penalized Wasserstein gradient flow 33
D.3 The velocity contribution from CVaR-penalization 34
D.4 Using CVaR vs smoothed CVaR for penalization 36
D.5 Proofs for the results concerning CVaR-penalized Wasserstein gradient flows with Clarke's generalized subgradients 37

A Related work

Generative models tailored for learning heavy-tailed targets There are many generative model designs tailored for heavy-tailed distributions (Allouche et al., 2022; 2026; Bhatia et al., 2021; Guan et al., 2026; Hickling & Prangle, 2025; Huster et al., 2021; Liu et al., 2026; Pandey et al., 2025). Even though tailored for heavy-tailed distributions, these generative models still exhibit several limitations. On the one hand, certain architectures (including mirror flow matching (Guan et al., 2026), Pareto GAN (Huster et al., 2021), Score-based Heavy-tailed Diffusion (Liu et al., 2026), Tail Transform Flow (TTF) (Hickling & Prangle, 2025), t-EDM (Pandey et al., 2025), and t-Flow (Pandey et al., 2025)) require prior knowledge or accurate estimates of the tail decay rate of the target distribution. Accurate estimation of the tail decay rate is itself a computationally challenging problem. On the other hand, many models, including EV-GAN (Allouche et al., 2022), ExceedGAN (Allouche et al., 2026), and ExGAN (Bhatia et al., 2021), operate over a fixed time horizon. As a result, the effective depth of the neural architecture is treated as a static hyperparameter determined a priori; this potentially limits the model's capacity to transport light-tailed sources to heavy-tailed targets. Lip-KL-GPA, a particle algorithm introduced in Gu et al. (2024) based on Lipschitz-regularized Wasserstein gradient flows, is both tail-agnostic and operates over an adaptive time horizon. As shown in Chen et al. (2025), Lip-KL-GPA and other generative models optimizing Lipschitz-regularized divergences outperform existing generative models on learning heavy-tailed targets, including f-GANs, optimal transport (OT) flow, continuous normalizing flows (CNFs), and score-based generative models (SGMs). However, the generative models optimizing Lipschitz-regularized divergences empirically exhibit premature saturation when learning heavy-tailed targets, the primary drawback that we address by introducing the CVaR penalization.

CVaR for tail-sensitive modeling CVaR is one instance of a broader family of spectral risk measures (Acerbi & Simonetti, 2002) that assign greater weight to the tail region. CVaR, also known as the Expected Shortfall, is a canonical choice in tail-sensitive modeling, with an extensive literature (Acerbi et al., 2001; Cont et al., 2025; Kishida, 2024), owing largely to its tractable variational representation via the Rockafellar–Uryasev formula (Rockafellar & Uryasev, 2002). Incorporating CVaR into the generative objective to improve the accuracy of learning heavy-tailed targets has also been explored by Tail-GAN (Cont et al., 2025), an adversarial generator for multi-asset financial returns that augments the GAN discriminator loss with Value-at-Risk (VaR) and CVaR, so that the generator learns to reproduce correct tail-risk statistics for benchmark financial portfolios. Unlike our method, however, its generator is a fixed-depth Lipschitz continuous transport map, which may structurally limit its ability to learn a heavy-tailed target distribution. Furthermore, the main goal of Tail-GAN is to reproduce tail statistics for financial portfolios, as opposed to our goal, which is learning the heavy-tailed target and capturing extreme events. A separate line of work incorporates CVaR into fine-tuning objectives for tail-sensitive reward maximization (Chaudhary et al., 2024; Santi et al., 2025; Wang et al., 2026). This line of research pursues a fundamentally different goal from ours: it reweights the generated distribution to favor extreme rewards, whereas we aim to learn heavy-tailed distributions and capture extreme events. The framework of Wang et al. (2026) is closest to ours, where CVaR, its Rockafellar–Uryasev representation, and its first variational derivative are employed for tail-sensitive reward maximization. However, the variational derivative of CVaR utilized in Santi et al. (2025); Wang et al. (2026) is computed under density assumptions that do not hold for general distributions such as empirical measures supported on finite samples. This necessitates either the derivation of the first variational subgradients of $\mathrm{CVaR}_\alpha^{Q,g}$, or the use of a smoothed CVaR, whose variational derivative exists for all $Q \in \mathcal{Q}^g$.

B Supplementary material on the smoothed CVaR and CVaR-GPA

B.1 Choice of the level α

The velocity component $\hat{v}_{Q,h}$ from the CVaR penalty term in Equation 8 acts, through the weight $\sigma\big((\|x\| - q_{\alpha,h}^{Q,g})/h\big)$, essentially only on the particles with $\|x\| > q_{\alpha,h}^{Q,g}$, since $\sigma\big((\|x\| - q_{\alpha,h}^{Q,g})/h\big)$ is the smooth surrogate of the indicator function $\mathbb{1}_{\{\|x\| > q_{\alpha,h}^{Q,g}\}}$. The level α needs to be extreme for two reasons.

First, there are infinitely many distributions Q for which $\mathrm{CVaR}_{\alpha,h}^{Q,g} = \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}$. As we observe in Figure 3, for moderate levels of α since the domain for CVaR is larger than that considered by an extreme level α, the CVaR discrepancy $\Delta C_h(Q; P^{\mathrm{tar}})$ nears 0, while Q is not close to P^{tar} in distribution. As $\alpha \uparrow 1$, the averaging region shrinks to the extreme tail, and the constraint $\Delta C_h(Q; P^{\mathrm{tar}}) = 0$ becomes a better constraint in capturing the distribution P^{tar}. Second, the goal of the velocity component $\hat{v}_{Q,h}$ is to place additional weight on the under-sampled tail region, to compensate for the vanishing velocity $-\nabla_x \phi^*$ from the Lipschitz-regularized KL divergence in that region. For a moderate level α, $\hat{v}_{Q,h}$ displaces a large number of moderate-radius particles by comparable amounts, and thereby perturbs the part of the distribution that D_{KL}^L already fits. For an extreme α, it acts only on the few particles that constitute the extreme tail, and leaves the bulk of the distribution to be learned by the Lipschitz-regularized KL divergence.

Table 3 and Figure 3 illustrate this phenomenon on the 2-dimensional Cauchy target, observed empirically when fine-tuning the Lip-KL-GPA pre-trained model at $\alpha \in \{0.85, 0.95, 0.9994\}$, that is, with $(1-\alpha)M \in \{750, 250, 3\}$ generated particles above the quantile, and with all other settings as in Section 4. At the two moderate levels, fine-tuning reduces the global error $\mathcal{E}_{L^1}$ only from 7.13 to 6.11 and 6.01, respectively, whereas at $\alpha = 0.9994$ it reduces it to 1.08. The tail error $\mathcal{E}_{\mathrm{tail}}$ decreases at all three levels, but the extreme level attains 0.0628, roughly five times smaller than the values 0.298 and 0.331 obtained at the moderate levels.

Table 3: 2-dimensional Cauchy target, Lip-KL-GPA pre-trained model: joint errors for three levels α, with $k = (1-\alpha)M$ the number of generated particles above the level.

	$\mathcal{E}_{L^1}$		$\mathcal{E}_{\mathrm{tail}}$	
α	pre-trained	fine-tuned	pre-trained	fine-tuned
0.85 ($k = 750$)	7.13	6.11	4.24	0.298
0.95 ($k = 250$)	7.13	6.01	4.24	0.331
0.9994 ($k = 3$)	7.13	1.08	4.24	0.0628

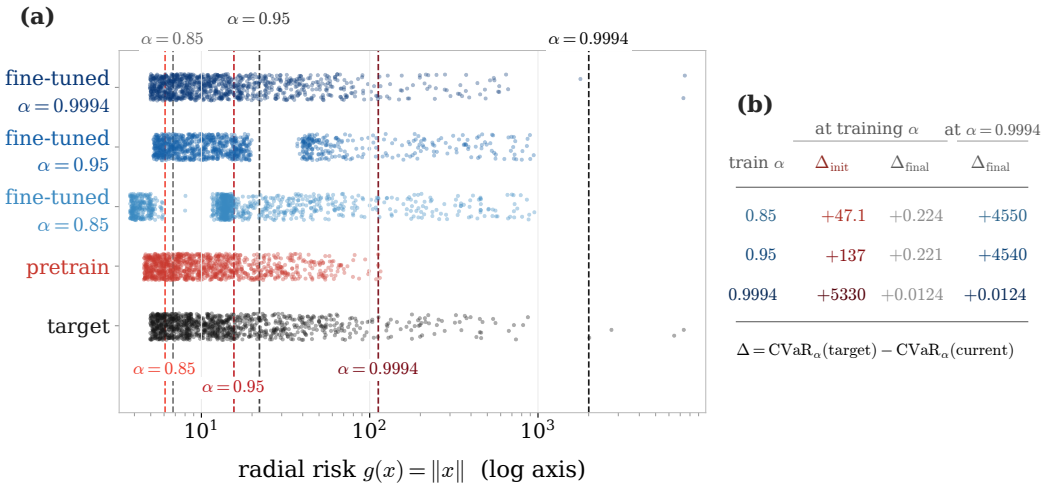

Figure 3: 2-dimensional Cauchy target. (a) The 1000 largest radii $g(x) = \|x\|$ of the target (black), of the pre-trained model (red), and of the fine-tuned models at three levels (blue). Dashed lines: VaR_α of the target (labels on top) and of the pre-trained model (labels below). (b) Smoothed CVaR discrepancy $\Delta C_h(Q; P^{\mathrm{tar}}) = \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g} - \mathrm{CVaR}_{\alpha,h}^{Q,g}$ before and after fine-tuning at the training level α, and after fine-tuning evaluated at $\alpha = 0.9994$.

B.2 Error metrics

For an empirical measure $\widehat{Q}_N$ with samples $\{Z^{(i)}\}_{i=1}^N$, the CDF and CCDF (complementary CDF) of the risk function g are given by $\Psi^{\widehat{Q}_N,g}(r) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{g(Z^{(i)})\le r\}}$ and $\bar{\Psi}^{\widehat{Q}_N,g} = 1 - \Psi^{\widehat{Q}_N,g}$, respectively. The **global L^1 error** is

$$\mathcal{E}_{L^1}(\widehat{Q}_N; \widehat{P^{\mathrm{tar}}}_N) = \int_0^{r_{\max}} \left|\bar{\Psi}^{\widehat{Q}_N,g}(r) - \bar{\Psi}^{\widehat{P^{\mathrm{tar}}}_N,g}(r)\right| dr, \tag{11}$$

where $r_{\max}$ is the largest risk value in either sample and the integral is evaluated by the trapezoidal rule on a uniform 2000-point grid of $[0, r_{\max}]$. Since $\bar{\Psi}$ is itself small in the tail, absolute CCDF differences are uninformative there, so we measure the **tail error** by the log-CCDF discrepancy over the tail quantiles,

$$\mathcal{E}_{\text{tail}}(\widehat{Q}_N; \widehat{P^{\mathrm{tar}}}_N) = \frac{1}{20}\sum_{j=1}^{20}\left|\log\frac{\bar{\Psi}^{\widehat{Q}_N,g}(t_j)+\varepsilon}{\bar{\Psi}^{\widehat{P^{\mathrm{tar}}}_N,g}(t_j)+\varepsilon}\right|, \qquad t_j = (\Psi^{\widehat{P^{\mathrm{tar}}}_N,g})^{-1}(p_j), \tag{12}$$

where $p_1, \dots, p_{20}$ are equi-spaced in $[0.95, 0.999]$ and $\varepsilon = 10^{-10}$ is added for numerical stability.

B.3 Algorithms

Algorithm 1 computes the smoothed empirical tail statistics used by CVaR-GPA, and Algorithm 2 states CVaR-GPA in full.

Algorithm 1 Smoothed empirical tail statistics

Require: samples $\{Z_i\}_{i=1}^m$, $g(\cdot) = \|\cdot\|$: the risk function, α: level, h: smoothing parameter, τ: root-finding tolerance

1: $g_i \leftarrow g(Z_i), \quad i = 1, \dots, m$

2: $q_h \leftarrow$ root of $q \mapsto \frac{1}{m}\sum_{i=1}^m \sigma\big((g_i - q)/h\big) - (1-\alpha)$ on the bracket Equation 18, to tolerance τ ▷ bisection; unique root by Equation 14 and Proposition B.4

3: $\mathrm{CVaR}_h \leftarrow q_h + \dfrac{1}{(1-\alpha)\,m}\displaystyle\sum_{i=1}^m r_h(g_i - q_h)$ ▷ empirical form of Equation 5

4: **return** $(q_h,\ \mathrm{CVaR}_h)$ ▷ $q_h = q_h^{\widehat{Q},g}$ and $\mathrm{CVaR}_h = \mathrm{CVaR}_{\alpha,h}^{\widehat{Q},g}$ for the empirical measure $\widehat{Q}$ of $\{Z_i\}$

B.4 Proofs of the results concerning the smoothed CVaR

B.4.1 Standing assumption and admissible perturbations

Throughout, $g : \mathbb{R}^d \to [0, \infty)$ is a measurable risk function; no Lipschitz continuity of g is assumed for the results on the smoothed CVaR. Throughout the appendix, we use the notation $\mathcal{F}_h$ to denote the loss functional penalized with smoothed CVaR, and $\mathcal{F}$ to denote the functional penalized with CVaR. The following assumption ensures that $D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})$, the CVaR discrepancies, and the loss functionals $\mathcal{F}_h$ and $\mathcal{F}$ are finite.

Assumption B.1. *The target distribution satisfies $P^{\mathrm{tar}} \in \mathcal{Q}^g$, and the generated distribution $Q \in \mathcal{Q}^g$ has a finite first moment.*

The condition $P^{\mathrm{tar}} \in \mathcal{Q}^g$ holds for every empirical target measure $\widehat{P^{\mathrm{tar}}}$, which is the measure used by CVaR-GPA. The condition $P^{\mathrm{tar}} \in \mathcal{Q}^g$ holds for every empirical target measure $\widehat{P^{\mathrm{tar}}}$, which is the measure used by CVaR-GPA. Thus CVaR-GPA can be applied even on extremely heavy-tailed targets with no finite first moments.

Variational derivatives and subgradients are taken with respect to perturbations of Q of the following form.

Algorithm 2 CVaR-GPA (smoothed CVaR)

Require: Hyperparameters α, λ, h, L, Δt; risk function $g(\cdot) = \|\cdot\|$; number of pre-trained particles M; number of target particles N; kinetic energy threshold ϵ; neural network approximation Γ_L^{NN}; number of network updates per iteration N_ϕ.

Result: Fine-tuned particles $\{Y_k^{(i)}\}_{i=1}^M$ at termination.

Sample $\{X^{(j)}\}_{j=1}^N \sim P^{\mathrm{tar}}$, a batch of samples from the target distribution.
Sample $\{Y_0^{(i)}\}_{i=1}^M \sim P^{\mathrm{pre}}$, a batch of samples from the pre-trained distribution.
Compute the target tail statistic $\mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g}$ via Algorithm 1 using $\{X^{(j)}\}_{j=1}^N$.
$k \leftarrow 0,\ \mathcal{K} \leftarrow 2\epsilon$
while $\mathcal{K} > \epsilon$ **do**
  $\phi_k^* \leftarrow \arg\max_{\phi \in \Gamma_L^{\mathrm{NN}}} \left\{ \frac{1}{M}\sum_i \phi(Y_k^{(i)}) - \log\left(\frac{1}{N}\sum_j e^{\phi(X^{(j)})}\right)\right\}$ ▷ N_ϕ steps
  $(q_{\alpha,h}^{\widehat{Q}_k,g}, \mathrm{CVaR}_{\alpha,h}^{\widehat{Q}_k,g})$ via Algorithm 1 using $\{Y_k^{(i)}\}_{i=1}^M$.
  $\Delta_k \leftarrow \mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{\widehat{Q}_k,g}$
  for $i = 1, \dots, M$ **do**
    $v_k^{(i)} \leftarrow -\nabla_x \phi_k^*(Y_k^{(i)}) + \frac{2\lambda}{1-\alpha}\,\Delta_k\,\sigma\Big(\frac{\|Y_k^{(i)}\| - q_{\alpha,h}^{\widehat{Q}_k,g}}{h}\Big)\,\frac{Y_k^{(i)}}{\|Y_k^{(i)}\|}$ ▷ Equation 10
  end for
  $Y_{k+1}^{(i)} \leftarrow Y_k^{(i)} + \Delta t\, v_k^{(i)}; \quad \mathcal{K} \leftarrow \frac{1}{2M}\sum_i \|v_k^{(i)}\|^2$
  $k \leftarrow k+1$
end while
return $\{Y_k^{(i)}\}_{i=1}^M$

Definition B.2 (Admissible perturbations)**.** *Let $Q \in \mathcal{Q}^g$. A signed measure ρ on $\mathbb{R}^d$ is a* right-admissible *perturbation at Q if*

$$\int_{\mathbb{R}^d} d\rho = 0, \qquad \|\rho\| := \int_{\mathbb{R}^d} d|\rho| < \infty, \qquad \int_{\mathbb{R}^d} \|x\|\, d|\rho| < \infty, \qquad \int_{\mathbb{R}^d} g\, d|\rho| < \infty.$$

It is left-admissible *if $-\rho$ is right-admissible, and* admissible *if it is both right- and left-admissible.*

Write $Q^\epsilon := Q + \epsilon\rho \in \mathcal{P}(\mathbb{R}^d)$ for all sufficiently small $\epsilon > 0$. For every right- (respectively left-) admissible ρ at Q and every ϵ with $Q^\epsilon \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathbb{E}_{Q^\epsilon}[g] \;\le\; \mathbb{E}_Q[g] + |\epsilon| \int g\, d|\rho| \;<\; \infty, \tag{13}$$

so that $Q^\epsilon \in \mathcal{Q}^g$. The condition $\int \|x\|\, d|\rho| < \infty$ ensures that Q^ϵ has a finite first moment whenever Q does.

B.4.2 The divergence property

The following proposition establishes that the functional $\mathcal{F}_h(Q; P^{\mathrm{tar}})$ is a divergence, so that the target distribution P^{tar} is its unique global minimizer.

Proposition B.3 (Divergence property)**.** *Let Assumption B.1 hold. Then, the loss functional $\mathcal{F}_h$ defined in Equation 1 satisfies $\mathcal{F}_h(Q; P^{\mathrm{tar}}) \ge 0$ for all $Q \in \mathcal{Q}^g$ and we have $\mathcal{F}_h(Q; P^{\mathrm{tar}}) = 0$ if and only if $Q = P^{\mathrm{tar}}$. That is, P^{tar} is the unique global minimizer of the optimization problem*

$$\inf_{Q \in \mathcal{Q}^g} \mathcal{F}_h(Q; P^{\mathrm{tar}}).$$

Proof. The Lipschitz-regularized KL divergence $D_{KL}^L(Q\|P^{\mathrm{tar}})$ satisfies the divergence property, i.e., $D_{KL}^L(Q\|P^{\mathrm{tar}}) \ge 0$ for all $Q \in \mathcal{P}(\mathbb{R}^d)$ and $D_{KL}^L(Q\|P^{\mathrm{tar}}) = 0$ if and only if $Q = P^{\mathrm{tar}}$ (see Birrell et al. (2022, Theorem 8) for the proof). This, combined with the fact that $(\Delta C_h(Q; P^{\mathrm{tar}}))^2 \ge 0$ for all $Q \in \mathcal{Q}^g$, proves that $\mathcal{F}_h(Q; P^{\mathrm{tar}}) \ge 0$ for all $Q \in \mathcal{Q}^g$ and $\mathcal{F}_h(Q; P^{\mathrm{tar}}) = 0$ if and only if $Q = P^{\mathrm{tar}}$. That is, P^{tar} is the unique global minimizer of the optimization problem $\inf_{Q \in \mathcal{Q}^g} \mathcal{F}_h(Q; P^{\mathrm{tar}})$. ■

B.4.3 PROPERTIES OF THE SMOOTHED CVAR

Proposition B.4 (Properties of the smoothed CVaR)**.** *Let $h > 0$, let $g : \mathbb{R}^d \to [0, \infty)$ be measurable, and let $Q \in \mathcal{Q}^g$. Then the function $F^{Q,g}_{\alpha,h}(\cdot) : \mathbb{R} \to \mathbb{R}$ is in C^∞ and strictly convex. Thus, it has a unique minimizer $q^{Q,g}_{\alpha,h}$, characterized by*

$$\mathbb{E}_Q\Big[\sigma\Big(\frac{g - q^{Q,g}_{\alpha,h}}{h}\Big)\Big] = 1 - \alpha, \quad \sigma(t) = \frac{1}{1 + e^{-t}}. \tag{14}$$

Furthermore, the variational derivative of the smoothed CVaR, with respect to admissible perturbations (Definition B.2), is given by

$$\frac{\delta}{\delta Q}\mathrm{CVaR}^{Q,g}_{\alpha,h} = \frac{1}{1-\alpha} r_h\left(g - q^{Q,g}_{\alpha,h}\right).$$

Proof. Let $h > 0$ be arbitrary. Since $0 \le r_h(g - y) \le g + |y| + h \log 2$ and $\mathbb{E}_Q[g] < \infty$, we have $\mathbb{E}_Q[r_h(g - y)] < \infty$ for all $y \in \mathbb{R}$. Thus it follows that $F^{Q,g}_{\alpha,h}(y)$ is finite for all $y \in \mathbb{R}$. Furthermore, for all $y \in \mathbb{R}$ we have

$$|\partial_y r_h(g - y)| = \sigma\left(\frac{g - y}{h}\right) \le 1.$$

Thus, by the dominated convergence theorem, we differentiate under the expectation in Equation 5 to obtain

$$\frac{d}{dy} F^{Q,g}_{\alpha,h}(y) = 1 - \frac{S^Q_h(y)}{1 - \alpha}, \tag{15}$$

where $S^Q_h(y) := \mathbb{E}_Q\Big[\sigma\Big(\frac{g-y}{h}\Big)\Big]$. Since $\sigma(\cdot) \in C^\infty$, and $\mathbb{E}_Q\Big[\sigma\Big(\frac{g-y}{h}\Big)\Big]$ is finite for all $y \in \mathbb{R}$, we have that $F^{Q,g}_{\alpha,h}(\cdot) \in C^\infty$. Furthermore, since $\sigma' > 0$ and $|\sigma'| \le \frac{1}{4}$, differentiating Equation 15 once more under the expectation gives $\frac{d^2}{dy^2} F^{Q,g}_{\alpha,h}(y) = \frac{1}{(1-\alpha)h}\mathbb{E}_Q\big[\sigma'\big(\frac{g-y}{h}\big)\big] > 0$, and thus $F^{Q,g}_{\alpha,h}(y)$ is strictly convex, and hence there is a unique minimizer $q^{Q,g}_{\alpha,h}$ such that

$$\frac{d}{dy} F^{Q,g}_{\alpha,h}(y)\Big|_{y = q^{Q,g}_{\alpha,h}} = 1 - \frac{S^Q_h(q^{Q,g}_{\alpha,h})}{1 - \alpha} = 0. \tag{16}$$

That is

$$\mathbb{E}_Q\Big[\sigma\Big(\frac{g - q^{Q,g}_{\alpha,h}}{h}\Big)\Big] = S^Q_h(q^{Q,g}_{\alpha,h}) = 1 - \alpha. \tag{17}$$

For empirical measures $\widehat{Q}$ with risk values $g_{(1)} \le \cdots \le g_{(m)}$ (order statistics), this characterization also yields the bracket used by the bisection in Algorithm 1: since $\sigma\big((g_{(m)} - q)/h\big) \le 1 - \alpha$ for $q \ge g_{(m)} + h \log \frac{\alpha}{1-\alpha}$ and $\sigma\big((g_{(1)} - q)/h\big) \ge 1 - \alpha$ for $q \le g_{(1)} + h \log \frac{\alpha}{1-\alpha}$,

$$q^{\widehat{Q},g}_{\alpha,h} \in \Big[g_{(1)} + h \log \frac{\alpha}{1 - \alpha},\ g_{(m)} + h \log \frac{\alpha}{1 - \alpha}\Big]. \tag{18}$$

To derive the variational derivative of $\mathrm{CVaR}^{Q,g}_{\alpha,h}$, we consider admissible perturbations ρ at Q in the sense of Definition B.2, and write $Q^\epsilon := Q + \epsilon\rho$. By Equation 13, $Q^\epsilon \in \mathcal{Q}^g$ for all ϵ with $Q^\epsilon \in \mathcal{P}(\mathbb{R}^d)$, so that $\mathrm{CVaR}^{Q^\epsilon,g}_{\alpha,h}$ is well defined.

Let ρ be an admissible perturbation at Q, set $q_0 := q^{Q,g}_{\alpha,h}$ and $q_\epsilon := q^{Q^\epsilon,g}_{\alpha,h}$. Define

$$R(y) := \frac{1}{1 - \alpha} \int r_h(g - y)\, d\rho. \tag{19}$$

Since $0 \le r_h(g - y) \le g + |y| + h \log 2$ and $\int g\, d|\rho| < \infty$ by admissibility, $R(y)$ is finite for every $y \in \mathbb{R}$. Since $r_h'(u) = \sigma(u/h) \le 1$, r_h is 1-Lipschitz continuous, and $R(\cdot)$ is $\frac{\|\rho\|}{1-\alpha}$-Lipschitz in y.

Since $F_{\alpha,h}^{Q^\epsilon,g} = F_{\alpha,h}^{Q,g} + \epsilon\, R$ by linearity of the expectation in Q, evaluating the infimum at q_0 and at q_ϵ respectively gives

$$\mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g} \le F_{\alpha,h}^{Q^\epsilon,g}(q_0) = \mathrm{CVaR}_{\alpha,h}^{Q,g} + \epsilon\, R(q_0), \tag{20}$$

$$\mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g} = F_{\alpha,h}^{Q,g}(q_\epsilon) + \epsilon\, R(q_\epsilon) \ge \mathrm{CVaR}_{\alpha,h}^{Q,g} + \epsilon\, R(q_\epsilon). \tag{21}$$

Hence, for $\epsilon > 0$, we have

$$R(q_\epsilon) \le \frac{1}{\epsilon}\left(\mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g} - \mathrm{CVaR}_{\alpha,h}^{Q,g}\right) \le R(q_0)$$

Similarly for $\epsilon < 0$, we have

$$R(q_0) \le \frac{1}{\epsilon}\left(\mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g} - \mathrm{CVaR}_{\alpha,h}^{Q,g}\right) \le R(q_\epsilon)$$

Fix $\eta > 0$ and set

$$\kappa := \min\left\{S_h^Q(q_0 - \eta) - (1-\alpha),\ (1-\alpha) - S_h^Q(q_0 + \eta)\right\}. \tag{22}$$

Since the mapping $S_h^Q(y)$ is strictly decreasing and equals $1-\alpha$ exactly at q_0, κ is positive.

Next, $Q^\epsilon = Q + \epsilon\rho$ gives, for every $y \in \mathbb{R}$,

$$S_h^{Q^\epsilon}(y) - S_h^Q(y) = \epsilon \int \sigma\Big(\frac{g-y}{h}\Big)\, d\rho, \tag{23}$$

and since $0 < \sigma < 1$,

$$\left|S_h^{Q^\epsilon}(y) - S_h^Q(y)\right| \le |\epsilon| \int \left|\sigma\Big(\frac{g-y}{h}\Big)\right| d|\rho| \le |\epsilon| \int d|\rho| = |\epsilon|\, \|\rho\|, \tag{24}$$

uniformly in y. Let now $|\epsilon|\,\|\rho\| < \kappa$. We have

$$S_h^{Q^\epsilon}(q_0 - \eta) \ge S_h^Q(q_0 - \eta) - |\epsilon|\,\|\rho\| \tag{25}$$

$$> S_h^Q(q_0 - \eta) - \kappa \tag{26}$$

$$\ge S_h^Q(q_0 - \eta) - \big(S_h^Q(q_0 - \eta) - (1-\alpha)\big) = 1 - \alpha. \tag{27}$$

Similarly we have

$$S_h^{Q^\epsilon}(q_0 + \eta) \le S_h^Q(q_0 + \eta) + |\epsilon|\,\|\rho\| \tag{28}$$

$$< S_h^Q(q_0 + \eta) + \kappa \tag{29}$$

$$\le S_h^Q(q_0 + \eta) + \big((1-\alpha) - S_h^Q(q_0 + \eta)\big) = 1 - \alpha. \tag{30}$$

Thus we have

$$S_h^{Q^\epsilon}(q_0 - \eta) > 1 - \alpha > S_h^{Q^\epsilon}(q_0 + \eta). \tag{31}$$

The minimizer q_ϵ is characterized by $S_h^{Q^\epsilon}(q_\epsilon) = 1-\alpha$, and $S_h^{Q^\epsilon}$ is strictly decreasing. If $q_\epsilon \le q_0 - \eta$, then

$$1 - \alpha = S_h^{Q^\epsilon}(q_\epsilon) \ \ge\ S_h^{Q^\epsilon}(q_0 - \eta) \ >\ 1 - \alpha, \tag{32}$$

a contradiction; and if $q_\epsilon \ge q_0 + \eta$, then

$$1 - \alpha = S_h^{Q^\epsilon}(q_\epsilon) \ \le\ S_h^{Q^\epsilon}(q_0 + \eta) \ <\ 1 - \alpha, \tag{33}$$

again a contradiction. Hence $|q_\epsilon - q_0| < \eta$ whenever $|\epsilon|\,\|\rho\| < \kappa$.

Since $\eta > 0$ was arbitrary and $\kappa = \kappa(\eta) > 0$, this says precisely that for every $\eta > 0$ there is $\epsilon_\eta > 0$ such that $|q_\epsilon - q_0| < \eta$ for all $|\epsilon| < \epsilon_\eta$; that is, $q_\epsilon \to q_0$ as $\epsilon \to 0$. Since R is Lipschitz, $R(q_\epsilon) \to R(q_0)$ as $\epsilon \to 0$, thus we have

$$\lim_{\epsilon \to 0} \frac{\mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g} - \mathrm{CVaR}_{\alpha,h}^{Q,g}}{\epsilon} = R(q_0) = \frac{1}{1-\alpha} \int r_h\big(g - q_{\alpha,h}^{Q,g}\big)\, d\rho. \tag{34}$$

The limit exists for every admissible ρ and is linear in ρ, so $\mathrm{CVaR}_{\alpha,h}^{\cdot,g}$ is Gateaux differentiable at Q with

$$\frac{\delta}{\delta Q}\mathrm{CVaR}_{\alpha,h}^{Q,g} = \frac{r_h\big(g - q_{\alpha,h}^{Q,g}\big)}{1-\alpha}. \tag{35}$$

■

B.4.4 PROOF OF THEOREM 2.2

Theorem 2.2. *Let $g \in C^1$ and $Q \in \mathcal{Q}^g$, and $h > 0$. The variational derivative of $\mathcal{F}_h$ is given by*

$$\frac{\delta \mathcal{F}_h(Q; P^{\mathrm{tar}})}{\delta Q} = \phi^* - \frac{2\lambda}{1-\alpha}\, \Delta C_h(Q; P^{\mathrm{tar}})\, r_h\big(g - q_{\alpha,h}^{Q,g}\big).$$

Proof. By Proposition B.4, $\mathrm{CVaR}_{\alpha,h}^{\cdot,g}$ is Gateaux differentiable at every $Q \in \mathcal{Q}^g$, with

$$\frac{\delta}{\delta Q}\mathrm{CVaR}_{\alpha,h}^{Q,g} = \frac{1}{1-\alpha}\, r_h\big(g - q_{\alpha,h}^{Q,g}\big).$$

Let ρ be an admissible perturbation at Q and write $Q^\epsilon = Q + \epsilon\rho$. Since $\Delta C_h(Q; P^{\mathrm{tar}}) = \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g} - \mathrm{CVaR}_{\alpha,h}^{Q,g}$ and the first term does not depend on Q,

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{\big(\Delta C_h(Q^\epsilon; P^{\mathrm{tar}})\big)^2 - \big(\Delta C_h(Q; P^{\mathrm{tar}})\big)^2}{\epsilon} \\
&= \lim_{\epsilon\to 0} \Big[\Delta C_h(Q^\epsilon; P^{\mathrm{tar}}) + \Delta C_h(Q; P^{\mathrm{tar}})\Big] \frac{\mathrm{CVaR}_{\alpha,h}^{Q,g} - \mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g}}{\epsilon} \\
&= -\,2\,\Delta C_h(Q; P^{\mathrm{tar}})\, \frac{1}{1-\alpha} \int r_h\big(g - q_{\alpha,h}^{Q,g}\big)\, d\rho,
\end{aligned}$$

where the limit of the bracket is $2\Delta C_h(Q; P^{\mathrm{tar}})$ by the continuity of $\epsilon \mapsto \mathrm{CVaR}_{\alpha,h}^{Q^\epsilon,g}$, itself a consequence of the Gateaux differentiability above. Hence

$$\frac{\delta}{\delta Q}\big(\Delta C_h(Q; P^{\mathrm{tar}})\big)^2 = -\frac{2}{1-\alpha}\, \Delta C_h(Q; P^{\mathrm{tar}})\, r_h\big(g - q_{\alpha,h}^{Q,g}\big).$$

Combining this with $\frac{\delta}{\delta Q} D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}}) = \phi^*$, and with the definition Equation 1 of $\mathcal{F}_h$, the linearity of the first variation gives

$$\frac{\delta \mathcal{F}_h(Q; P^{\mathrm{tar}})}{\delta Q} = \phi^* - \frac{2\lambda}{1-\alpha}\, \Delta C_h(Q; P^{\mathrm{tar}})\, r_h\big(g - q_{\alpha,h}^{Q,g}\big),$$

which is the asserted first variation. Both terms are finite under Assumption B.1: $D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})$ and ϕ^* by Chen et al. (2025, Theorem 2, Theorem 3), and the CVaR term by Proposition B.4.

The smoothed CVaR-penalized Wasserstein gradient flow is Equation 6 with $\mathcal{F} = \mathcal{F}_h$ and $Q_0 = P^{\mathrm{pre}}$, with velocity field $v_{Q,h} = -\nabla_x \frac{\delta \mathcal{F}_h(Q;P^{\mathrm{tar}})}{\delta Q}$. Since $r_h'(u) = \sigma(u/h)$ and $g \in C^1$, the chain rule gives, for every $x \in \mathbb{R}^d$,

$$\nabla_x\, r_h\big(g(x) - q_{\alpha,h}^{Q,g}\big) = \sigma\Big(\frac{g(x) - q_{\alpha,h}^{Q,g}}{h}\Big)\nabla_x g(x),$$

and therefore

$$v_{Q,h}(x) = -\nabla_x \phi^*(x) + \frac{2\lambda}{1-\alpha}\, \Delta C_h(Q; P^{\mathrm{tar}})\, \sigma\Big(\frac{g(x) - q_{\alpha,h}^{Q,g}}{h}\Big)\nabla_x g(x),$$

which is Equation 8. Here $\nabla_x \phi^*$ exists almost everywhere by Rademacher's theorem, since $\phi^* \in \Gamma_L$ is Lipschitz continuous. ∎

B.4.5 DIMENSION-FREE ESTIMATION OF CVAR-PENALIZATION

The following lemma is the precise form of Lemma 2.1.

Lemma B.5 (Dimension-free estimation of CVaR-penalization). *Let $\{X^{(j)}\}_{j=1}^N$ be i.i.d. samples of P^{tar}, and let $\widehat{P^{\mathrm{tar}}} := \frac{1}{N}\sum_{j=1}^N \delta_{X^{(j)}}$ denote the corresponding empirical measure. Assume that $\mathbb{E}_{P^{\mathrm{tar}}}[g^q] < \infty$ for some $q > 1$. Then there exists a constant C, depending only on q and $\mathbb{E}_{P^{\mathrm{tar}}}[g^q]$ and independent of the dimension d and the smoothing parameter $h > 0$, such that*

$$\mathbb{E}_{(P^{\mathrm{tar}})^{\otimes N}}\left|\mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}\right| \le \frac{C}{1-\alpha}\Big(N^{-\frac{1}{2}} + N^{-\frac{q-1}{q}}\Big). \tag{36}$$

If $q = 2$, the bound holds with q replaced by any $q' \in (1, 2)$.

Proof. Let $Q_1, Q_2 \in \mathcal{Q}^g$ and $\mu_i := g_{\#}Q_i$ for $i = 1, 2$. For every Borel measurable function $\psi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_{Q_i}[\psi \circ g]$ is finite, we have from the change of variables formula

$$\int_{\mathbb{R}^d} \psi\big(g(x)\big)\, dQ_i(x) = \int_{\mathbb{R}} \psi(z)\, d\mu_i(z). \tag{37}$$

Taking $\psi(z) = |z|$ and using $Q_i \in \mathcal{Q}^g$ shows that μ_i has a finite first moment and thus $\mathcal{W}_1(\mu_1, \mu_2) < \infty$. Taking $\psi(z) = r_h(z - y)$, and recalling that $0 \le r_h(g - y) \le g + |y| + h\log 2$ and $\mathbb{E}_{Q_i}[g] < \infty$ imply $\mathbb{E}_{Q_i}[r_h(g - y)] = \mathbb{E}_{Q_i}[\psi \circ g] < \infty$, we may rewrite the smoothed Rockafellar–Uryasev objective as

$$F_{\alpha,h}^{Q_i,g}(y) = y + \frac{1}{1-\alpha}\int_{\mathbb{R}} r_h(z - y)\, d\mu_i(z). \tag{38}$$

For each fixed $y \in \mathbb{R}$, the map $z \mapsto r_h(z - y)$ is 1-Lipschitz. Hence we have

$$\big|F_{\alpha,h}^{Q_1,g}(y) - F_{\alpha,h}^{Q_2,g}(y)\big| = \frac{1}{1-\alpha}\left|\int_{\mathbb{R}} r_h(z - y)\, d\big[\mu_1 - \mu_2\big](z)\right| \le \frac{1}{1-\alpha}\,\mathcal{W}_1(\mu_1, \mu_2), \tag{39}$$

for every $y \in \mathbb{R}$, where the last inequality follows from the Kantorovich–Rubinstein duality (Kantorovich & Rubinstein, 1958).

Furthermore, we have

$$\begin{aligned}
\mathrm{CVaR}_{\alpha,h}^{Q_1,g} &\le F_{\alpha,h}^{Q_1,g}\big(q_{\alpha,h}^{Q_2,g}\big) && (40)\\
&\le F_{\alpha,h}^{Q_2,g}\big(q_{\alpha,h}^{Q_2,g}\big) + \big|F_{\alpha,h}^{Q_1,g}\big(q_{\alpha,h}^{Q_2,g}\big) - F_{\alpha,h}^{Q_2,g}\big(q_{\alpha,h}^{Q_2,g}\big)\big| && (41)\\
&\le F_{\alpha,h}^{Q_2,g}\big(q_{\alpha,h}^{Q_2,g}\big) + \frac{1}{1-\alpha}\,\mathcal{W}_1(\mu_1, \mu_2) && (42)\\
&= \mathrm{CVaR}_{\alpha,h}^{Q_2,g} + \frac{1}{1-\alpha}\,\mathcal{W}_1(\mu_1, \mu_2). && (43)
\end{aligned}$$

Similarly, we have

$$\mathrm{CVaR}_{\alpha,h}^{Q_2,g} \le \mathrm{CVaR}_{\alpha,h}^{Q_1,g} + \frac{1}{1-\alpha}\,\mathcal{W}_1(\mu_1, \mu_2). \tag{44}$$

Combining Equation 40 and Equation 44, we obtain

$$\big|\mathrm{CVaR}_{\alpha,h}^{Q_1,g} - \mathrm{CVaR}_{\alpha,h}^{Q_2,g}\big| \le \frac{1}{1-\alpha}\,\mathcal{W}_1(\mu_1, \mu_2) = \frac{1}{1-\alpha}\,\mathcal{W}_1(g_{\#}Q_1, g_{\#}Q_2). \tag{45}$$

Thus we have

$$\left|\mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}\right| \le \frac{1}{1-\alpha}\,\mathcal{W}_1(g_{\#}P^{\mathrm{tar}}, g_{\#}\widehat{P^{\mathrm{tar}}}) \tag{46}$$

From Fournier & Guillin (2015, Theorem 1), we have that, for a measure $\mu \in \mathcal{P}(\mathbb{R})$, if there exists $q > 1$ and $q \ne 2$ such that μ has finite q^{th} moment,

$$\mathbb{E}_{(\mu)^{\otimes N}}\Big|\mathcal{W}_1(\mu, \widehat{\mu}_N)\Big| \le C(N^{-\frac{1}{2}} + N^{-\frac{(q-1)}{q}}), \tag{47}$$

where C is a coefficient that depends only on q and the qth moment of μ. Combining Equation 46 and Equation 47, we obtain

$$\mathbb{E}_{(P^{\mathrm{tar}})^{\otimes N}}\left|\mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{P^{\mathrm{tar}},g}\right| \le \frac{C}{1-\alpha}(N^{-\frac{1}{2}} + N^{-\frac{(q-1)}{q}}) \tag{48}$$

where C depends only on q and the qth moment of $g_{\#}P^{\mathrm{tar}}$. If $q = 2$, then $\mathbb{E}_{P^{\mathrm{tar}}}[g^{q'}] < \infty$ for every $q' \in (1, 2)$, and the above argument applies with q' in place of q. ∎

Table 4: Seven pre-trained generative models before and after fine-tuning with CVaR-GPA, all with the same hyperparameters. Factor: pre-trained error divided by fine-tuned error. For Fama–French 25 and streamflow, the pre-trained and fine-tuned errors are geometric means over the d marginals, so that the factor equals the geometric mean of the per-marginal error-reduction factors (up to rounding of the displayed errors). Bold: error-reduction factor above 1 (improvement). The tail parameter of t-EDM is selected by the tail error of the pre-trained model over $\nu \in \{3, 5, 7\}$ ($\nu = 3$ on the Cauchy and Fama–French targets, $\nu = 5$ on the streamflow data); on Neal's funnel it uses the per-coordinate setting $\nu = (4, 20)$ of its authors.

	$\mathcal{E}_{L^1}$			$\mathcal{E}_{\text{tail}}$		
pre-trained model	pre-trained	fine-tuned	factor	pre-trained	fine-tuned	factor
2-d Cauchy ($\nu = 1$), joint $\|x\|$, $M{=}N{=}5000$						
Lip-KL-GPA	7.13	1.08	**6.6**×	4.24	0.063	**67.3**×
TTF	17.6	7.13	**2.5**×	0.138	0.149	0.9×
Tail-GAN	50.5	1.53	**33.0**×	4.95	0.0379	**130.6**×
t-EDM	6.28	1.22	**5.1**×	0.294	0.0512	**5.7**×
SHD	278	51.2	**5.4**×	2.8	1.02	**2.7**×
DLPM	604	17.3	**35.0**×	0.659	0.419	**1.6**×
mTAF	67.6	14.0	**4.8**×	1.81	0.347	**5.2**×
2-d Neal's funnel, joint $\|x\|$, $M{=}N{=}5000$						
Lip-KL-GPA	0.102	0.0229	**4.5**×	0.0763	0.0166	**4.6**×
TTF	0.524	0.0385	**13.6**×	0.276	0.0172	**16.0**×
Tail-GAN	0.839	0.025	**33.6**×	2.45	0.0316	**77.5**×
t-EDM	0.192	0.0246	**7.8**×	0.0872	0.0141	**6.2**×
SHD	1.74	0.0235	**73.9**×	0.596	0.0313	**19.0**×
DLPM	8.15	0.0305	**267.4**×	0.237	0.0226	**10.5**×
mTAF	0.565	0.0228	**24.8**×	0.429	0.0207	**20.7**×
Fama–French 25, per-portfolio geometric mean, $M{=}N{=}1182$						
Lip-KL-GPA	0.529	0.125	**4.2**×	3.89	0.108	**36.1**×
TTF	5.64	0.173	**32.7**×	0.848	0.138	**6.2**×
Tail-GAN	1.33	0.138	**9.6**×	2.21	0.11	**20.1**×
t-EDM	0.426	0.123	**3.5**×	0.427	0.103	**4.1**×
SHD	5.16	0.152	**34.0**×	2.34	0.113	**20.7**×
DLPM	58.4	0.125	**468.2**×	2.92	0.116	**25.0**×
mTAF	1.69	0.142	**11.9**×	2.77	0.0963	**28.8**×
Streamflow (64 gauges), per-gauge geometric mean, $M{=}N{=}7305$						
Lip-KL-GPA	0.21	0.121	**1.7**×	1.74	0.151	**11.6**×
TTF	0.585	0.118	**4.9**×	0.226	0.146	**1.6**×
Tail-GAN	0.694	0.238	**2.9**×	0.921	0.202	**4.6**×
t-EDM	0.187	0.0848	**2.2**×	0.267	0.0886	**3.0**×
SHD	18.4	0.192	**95.9**×	2.63	0.195	**13.5**×
DLPM	277	0.414	**667.7**×	3.94	0.371	**10.6**×
mTAF	0.209	0.107	**2.0**×	0.704	0.179	**3.9**×

C Numerical Experiments

For the 2-dimensional synthetic targets we report the error reduction factors on their joint distribution, and for the high-dimensional Fama–French 25 and streamflow targets, the factor reported for each model is itself the geometric mean of the per-marginal error-reduction factors over the d marginals.

C.1 Full tables of errors corresponding to Table 1

Table 4 reports the errors underlying Table 1, and Table 5 reports the corresponding results for the Student-t targets with tail indices $\nu \in \{1, 1.2, 1.5, 1.8\}$.

Table 5: 2-dimensional isotropic Student-t targets with $\nu \in \{1, 1.2, 1.5, 1.8\}$: seven pre-trained generative models before and after fine-tuning with CVaR-GPA, all with the same hyperparameters. Factor: pre-trained error divided by fine-tuned error. Bold: error-reduction factor above 1 (improvement).

	$\mathcal{E}_{L^1}$			$\mathcal{E}_{\text{tail}}$		
pre-trained model	pre-trained	fine-tuned	factor	pre-trained	fine-tuned	factor
2-d Student-t ($\nu = 1$, Cauchy), joint $\|x\|$, $M{=}N{=}5000$						
Lip-KL-GPA	7.13	1.08	**6.6**×	4.24	0.063	**67.3**×
TTF	17.6	7.13	**2.5**×	0.138	0.149	0.9×
Tail-GAN	50.5	1.53	**33.0**×	4.95	0.0379	**130.6**×
t-EDM	6.28	1.22	**5.1**×	0.294	0.0512	**5.7**×
SHD	278	51.2	**5.4**×	2.8	1.02	**2.7**×
DLPM	604	17.3	**35.0**×	0.659	0.419	**1.6**×
mTAF	67.6	14.0	**4.8**×	1.81	0.347	**5.2**×
2-d Student-t ($\nu = 1.2$), joint $\|x\|$, $M{=}N{=}5000$						
Lip-KL-GPA	1.65	0.302	**5.5**×	2.06	0.0324	**63.6**×
TTF	223	0.375	**594.9**×	0.126	0.0661	**1.9**×
Tail-GAN	4.29	0.309	**13.9**×	14.3	0.0377	**378.1**×
t-EDM	0.872	0.0945	**9.2**×	0.137	0.0263	**5.2**×
SHD	5.74	1.91	**3.0**×	0.352	0.186	**1.9**×
DLPM	92.3	0.523	**176.5**×	0.321	0.0822	**3.9**×
mTAF	3.02	0.422	**7.1**×	0.937	0.0733	**12.8**×
2-d Student-t ($\nu = 1.5$), joint $\|x\|$, $M{=}N{=}5000$						
Lip-KL-GPA	0.381	0.0184	**20.7**×	0.981	0.017	**57.6**×
TTF	10.3	0.0711	**144.4**×	0.145	0.0395	**3.7**×
Tail-GAN	1.64	0.0527	**31.1**×	19.1	0.0431	**442.6**×
t-EDM	0.371	0.0374	**9.9**×	0.0856	0.0281	**3.0**×
SHD	15.0	10.4	**1.4**×	1.32	0.985	**1.3**×
DLPM	25.6	0.0744	**343.6**×	0.289	0.0247	**11.7**×
mTAF	0.642	0.0614	**10.5**×	0.483	0.0281	**17.2**×
2-d Student-t ($\nu = 1.8$), joint $\|x\|$, $M{=}N{=}5000$						
Lip-KL-GPA	0.106	0.0184	**5.8**×	0.06	0.0303	**2.0**×
TTF	0.339	0.0251	**13.5**×	0.221	0.0254	**8.7**×
Tail-GAN	0.378	0.0478	**7.9**×	2.2	0.0369	**59.8**×
t-EDM	0.22	0.0214	**10.3**×	0.107	0.0242	**4.4**×
SHD	1.44	0.0278	**51.7**×	0.477	0.0214	**22.2**×
DLPM	12.8	0.0466	**274.7**×	0.272	0.0381	**7.1**×
mTAF	0.257	0.0316	**8.1**×	0.242	0.015	**16.2**×

C.2 Student-t family

C.3 Neal's funnel distribution

Figure 4 compares, one pre-trained model at a time, the samples generated before and after fine-tuning with CVaR-GPA, together with the target samples and the target log-density used as the reference throughout. All panels use $M = N = 5000$ particles, the same hyperparameters, and the same axes. The corresponding per-marginal errors and error-reduction factors are reported in Table 8.

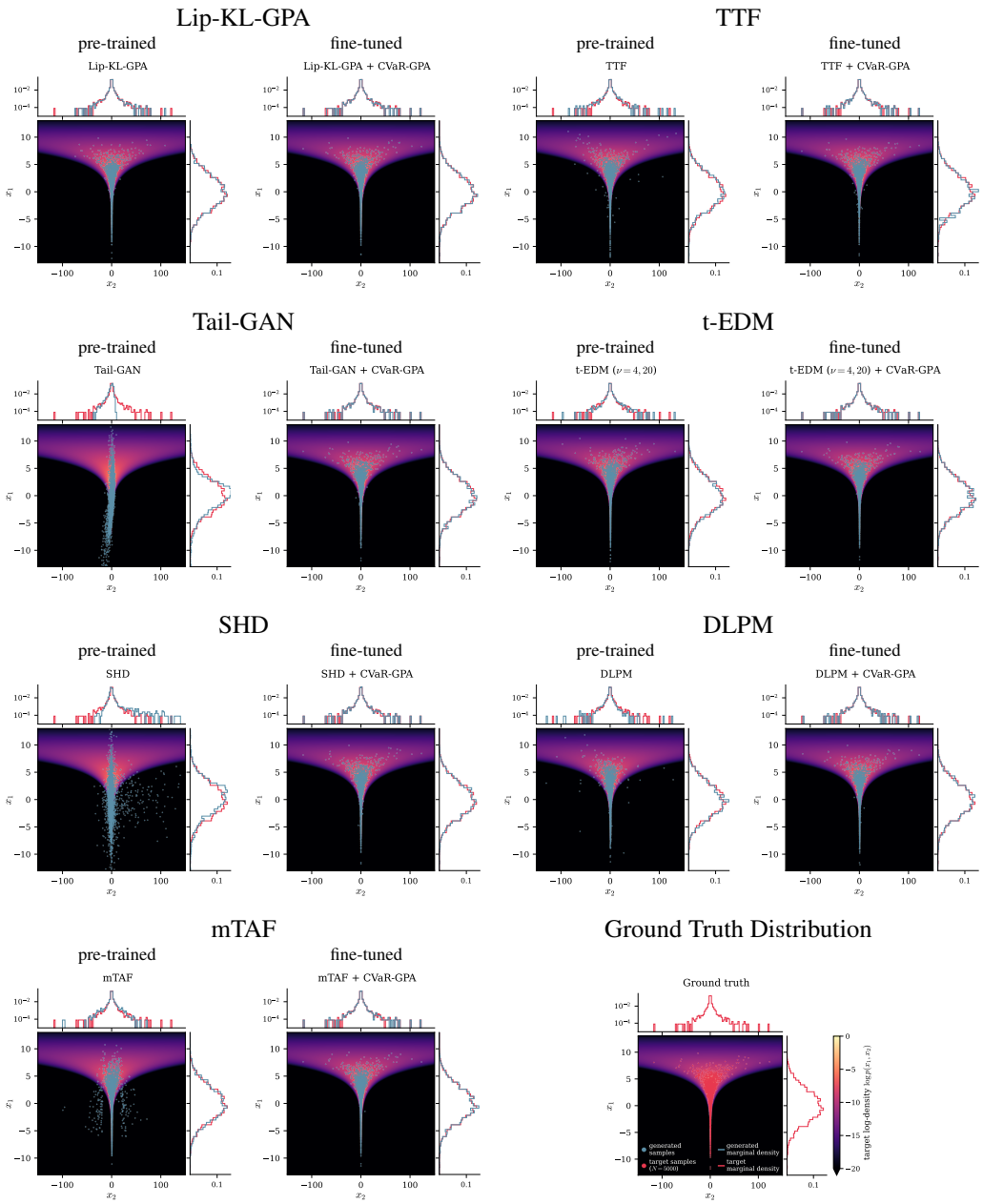

Figure 4: Neal's funnel, $M = N = 5000$. Each pair of columns shows one pre-trained model before and after fine-tuning with CVaR-GPA, under the same hyperparameters for all seven models.

Three features are common to the seven comparisons. First, the pre-trained models differ widely from one another, both in the accuracy of the joint structure and in overall scale, yet the fine-tuned models reach a comparable accuracy: the joint errors of the seven fine-tuned models lie between 0.0228 and 0.0385 ($\mathcal{E}_{L^1}$) and between 0.0141 and 0.0316 ($\mathcal{E}_{\text{tail}}$) (Table 4). Second, CVaR-GPA improves the heavy-tailed marginal x_2 and the Gaussian marginal x_1 simultaneously, even though the penalization acts on the joint radial statistic $g(x) = \|x\|$ and uses no per-coordinate information; this is the behavior expected of a method defined on the joint distribution rather than marginal by marginal, and it is what makes CVaR-GPA suitable for anisotropic targets with correlated marginals. Third, the largest gains occur where the pre-trained model is least accurate: DLPM, whose scale is wrong by orders of magnitude, improves by $389.7\times$ and $172.6\times$ in global error, while Tail-GAN improves by $27.2\times$ and $118.4\times$ in tail error (Table 8). The only entry that does not improve is the tail error of the Gaussian marginal x_1 for Lip-KL-GPA, which is essentially unchanged (0.0394 to 0.0403).

C.4 Per-marginal results

The joint errors in Table 4 are complemented here by the errors of the individual marginal distributions. Table 6 reports, for each dataset and pre-trained model, the number of marginals whose error decreases after fine-tuning; Tables 7 and 8 report the per-marginal errors for the 2-dimensional Cauchy target and Neal's funnel, and Figures 5 and 6 summarize the per-marginal errors for Fama–French 25 and streamflow, which are listed in full in Appendices C.5 and C.6.

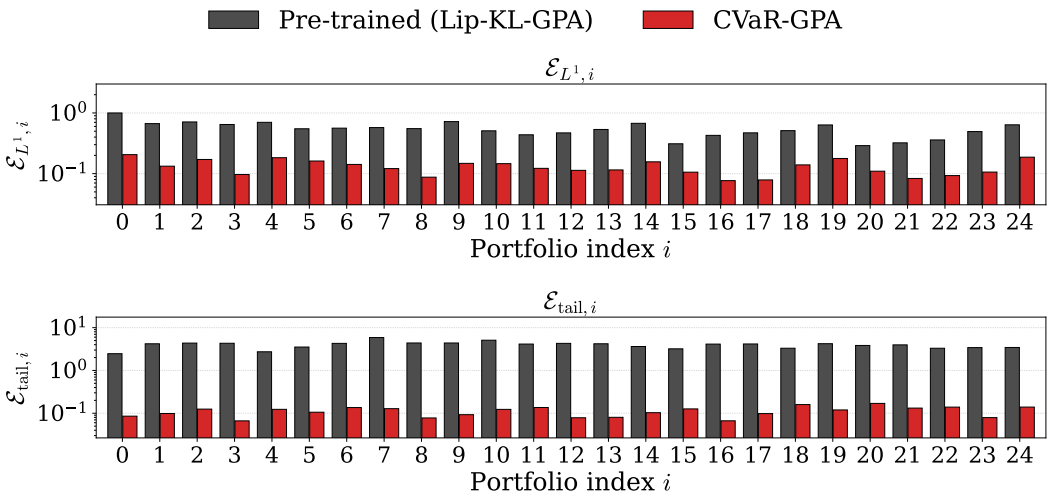

Figure 5: Fama–French 25: per-portfolio errors of the Lip-KL-GPA model before and after fine-tuning with CVaR-GPA.

C.5 Fama–French 25: per-portfolio results

Tables 9 and 10 list the global and the tail error of every portfolio for the seven pre-trained models and the fine-tuned models.

Table 9: Fama–French 25: $\mathcal{E}_{L^1}$ of each portfolio for the seven pre-trained models (pre.) and the fine-tuned models (f.t.).

	Lip-KL-GPA		TTF		Tail-GAN		t-EDM		SHD		DLPM		mTAF	
portfolio	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.
0	1	0.206	1.5	0.217	1.02	0.215	0.335	0.203	8.97	0.23	89.6	0.221	3.34	0.231
1	0.667	0.133	6.18	0.229	0.787	0.168	0.383	0.171	7.42	0.15	59.9	0.15	2.49	0.179
2	0.711	0.171	1.14	0.176	3.24	0.144	0.428	0.159	6.76	0.178	65.5	0.164	2.4	0.181
3	0.645	0.0967	4.05	0.19	2.58	0.153	0.507	0.125	5.42	0.181	62.2	0.131	1.93	0.17

	Lip-KL-GPA		TTF		Tail-GAN		t-EDM		SHD		DLPM		mTAF	
portfolio	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.
4	0.701	0.183	25.1	0.202	1.27	0.176	0.47	0.167	6.48	0.203	62.8	0.148	2.26	0.151
5	0.549	0.161	30.7	0.194	0.607	0.177	0.379	0.137	4.88	0.136	67.0	0.152	1.95	0.138
6	0.562	0.142	21.6	0.159	1.11	0.162	0.488	0.13	5.04	0.14	68.5	0.128	1.8	0.164
7	0.574	0.121	27.6	0.198	0.453	0.125	0.496	0.109	5.35	0.123	62.7	0.113	1.54	0.154
8	0.552	0.0874	0.714	0.164	3.86	0.122	0.571	0.127	5.21	0.182	53.2	0.106	1.59	0.116
9	0.72	0.148	2.86	0.227	0.637	0.153	0.629	0.109	5.87	0.183	74.3	0.116	2.29	0.189
10	0.508	0.146	1.02	0.148	3.15	0.138	0.284	0.0948	4.96	0.121	62.5	0.109	1.72	0.178
11	0.437	0.122	2.62	0.157	0.59	0.0884	0.393	0.103	4.23	0.129	53.2	0.0972	1.53	0.122
12	0.469	0.113	42.0	0.13	1.23	0.136	0.462	0.106	4.24	0.125	48.0	0.0883	1.2	0.143
13	0.535	0.115	41.0	0.198	1.94	0.121	0.462	0.112	4.99	0.143	46.0	0.12	1.53	0.11
14	0.675	0.156	2.47	0.173	2.62	0.172	0.507	0.142	6.45	0.135	64.0	0.155	2.1	0.134
15	0.31	0.106	2.59	0.0976	1.29	0.0989	0.238	0.109	4.09	0.089	40.9	0.0943	1.43	0.117
16	0.427	0.0766	113	0.139	1.38	0.114	0.334	0.0906	4.26	0.119	63.0	0.11	1.39	0.0935
17	0.47	0.0787	2.16	0.131	1.1	0.105	0.478	0.114	4.71	0.164	53.9	0.101	1.36	0.115
18	0.511	0.139	1.07	0.149	0.806	0.136	0.49	0.12	5.05	0.149	54.8	0.104	1.62	0.115
19	0.634	0.178	1.28	0.185	0.683	0.186	0.751	0.159	7.13	0.17	71.8	0.2	1.99	0.176
20	0.289	0.11	5.38	0.129	1.19	0.125	0.262	0.106	3.2	0.152	51.4	0.101	1.25	0.0931
21	0.322	0.0833	37.8	0.153	2	0.114	0.237	0.0722	3.14	0.124	38.6	0.148	0.896	0.112
22	0.359	0.0928	30.9	0.198	1.59	0.112	0.373	0.123	4.38	0.158	39.1	0.077	0.963	0.111
23	0.493	0.106	1.85	0.234	1.62	0.121	0.464	0.103	4.64	0.181	71.8	0.108	1.43	0.164
24	0.637	0.187	1.33	0.242	2.44	0.18	0.698	0.16	6.07	0.216	64.5	0.186	2.22	0.212

Table 10: Fama–French 25: $\mathcal{E}_{\text{tail}}$ of each portfolio for the seven pre-trained models (pre.) and the fine-tuned models (f.t.).

	Lip-KL-GPA		TTF		Tail-GAN		t-EDM		SHD		DLPM		mTAF	
portfolio	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.
0	2.47	0.0855	0.713	0.102	1.97	0.115	0.934	0.0608	2.61	0.12	2.89	0.08	2.32	0.0976
1	4.2	0.0987	0.936	0.192	3.04	0.0794	0.192	0.121	2.58	0.0893	2.87	0.0621	1.21	0.122
2	4.36	0.125	0.709	0.0715	1.87	0.121	0.211	0.102	3.32	0.121	2.86	0.16	3.12	0.112
3	4.32	0.0664	0.999	0.236	1.13	0.0915	0.335	0.0995	2.34	0.125	2.92	0.064	3.01	0.138
4	2.73	0.123	1.08	0.176	1.31	0.129	0.287	0.0628	3.17	0.135	2.89	0.0918	1.28	0.0903
5	3.53	0.106	0.86	0.129	0.352	0.136	0.287	0.0955	2.3	0.138	2.94	0.114	4.2	0.0716
6	4.3	0.136	0.835	0.141	2.23	0.0847	0.377	0.116	2.5	0.0829	2.93	0.0808	4.22	0.107
7	5.88	0.128	0.825	0.142	1.89	0.0781	0.349	0.0754	2.54	0.0841	2.97	0.0584	3.26	0.079
8	4.39	0.0776	1.26	0.12	1.46	0.0885	0.412	0.0777	2.51	0.0598	2.93	0.0329	2.34	0.0848
9	4.38	0.0926	0.902	0.199	1.89	0.0944	0.391	0.0868	2.43	0.139	2.91	0.0711	2.25	0.0729
10	5.09	0.123	0.682	0.131	1.22	0.0453	1.17	0.0959	1.62	0.0887	2.97	0.108	4.31	0.129
11	4.14	0.136	0.864	0.156	0.434	0.102	0.487	0.1	2.28	0.132	2.95	0.104	3.36	0.101
12	4.3	0.0784	0.717	0.124	1.6	0.142	0.475	0.0746	2.46	0.103	2.95	0.0643	3.91	0.06
13	4.2	0.0803	0.797	0.15	1.86	0.0778	0.376	0.0508	2.55	0.0666	2.91	0.102	2.3	0.0533
14	3.63	0.103	1.02	0.117	17.5	0.144	1.15	0.129	2.43	0.0955	2.92	0.118	2.23	0.0775
15	3.2	0.126	0.976	0.104	1.53	0.134	0.336	0.12	1.62	0.0826	2.96	0.92	3.49	0.0821
16	4.13	0.0666	0.644	0.0844	6.91	0.0694	0.43	0.09	1.68	0.113	2.98	0.0705	3.21	0.0698
17	4.16	0.0982	0.621	0.171	3.19	0.0978	0.421	0.0741	2.57	0.117	2.93	0.101	3.2	0.069
18	3.31	0.16	0.727	0.159	4.01	0.168	1.23	0.893	2.58	0.13	2.89	0.131	3.13	0.155
19	4.22	0.119	0.706	0.1	3.17	0.117	0.433	0.104	2.69	0.119	2.88	0.139	3.05	0.079
20	3.84	0.17	1.11	0.164	1.14	0.179	0.414	0.167	1.48	0.231	2.93	0.12	3.88	0.152
21	3.96	0.132	1.12	0.214	15.8	0.141	0.337	0.114	2.19	0.186	2.92	0.917	3.16	0.144
22	3.3	0.139	1.05	0.143	10.6	0.128	0.315	0.0926	2.58	0.135	2.87	0.869	2.03	0.113
23	3.42	0.0791	0.907	0.208	1.66	0.167	0.351	0.11	2.28	0.114	2.92	0.092	3.2	0.109
24	3.45	0.139	0.575	0.0751	1.69	0.162	0.434	0.12	2.29	0.14	2.86	0.0849	1.42	0.15

C.6 STREAMFLOW: PER-GAUGE RESULTS

The streamflow target consists of the daily specific discharge (mm/day) of the 64 gauges with the smallest drainage areas (2.5 to 204 km^2) among the 312 gauges of the Ohio River basin with a

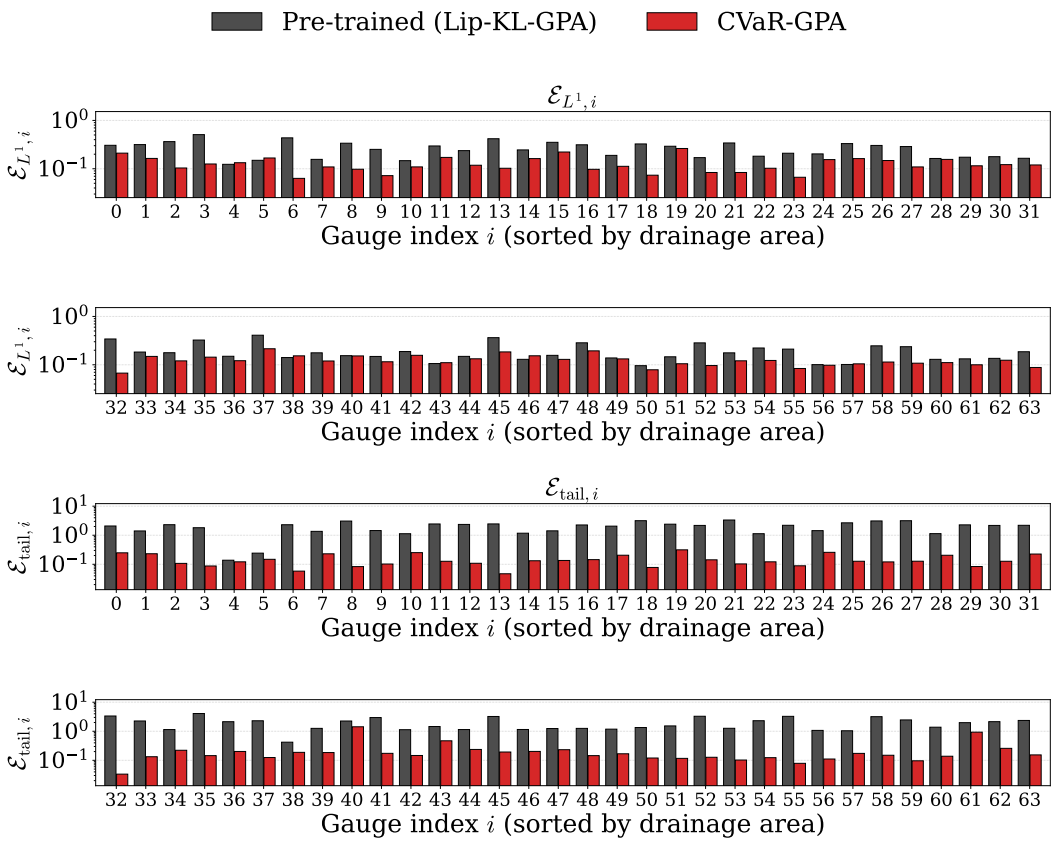

Figure 6: Streamflow: per-gauge errors of the Lip-KL-GPA model before and after fine-tuning with CVaR-GPA.

Table 6: Number of marginal distributions whose error decreases after fine-tuning with CVaR-GPA, out of the d marginals of each dataset, for the seven pre-trained models.

	2-d Cauchy ($d = 2$)		Neal's funnel ($d = 2$)		FF25 ($d = 25$)		Streamflow ($d = 64$)	
pre-trained model	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$	$\mathcal{E}_{L^1}$	$\mathcal{E}_{\text{tail}}$
Lip-KL-GPA	2/2	2/2	2/2	1/2	25/25	25/25	58/64	64/64
TTF	2/2	1/2	2/2	2/2	25/25	25/25	64/64	45/64
Tail-GAN	2/2	2/2	2/2	2/2	25/25	25/25	64/64	57/64
t-EDM	2/2	2/2	2/2	2/2	25/25	25/25	64/64	61/64
SHD	1/2	1/2	2/2	2/2	25/25	25/25	64/64	64/64
DLPM	2/2	2/2	2/2	2/2	25/25	25/25	64/64	64/64
mTAF	2/2	2/2	2/2	2/2	25/25	25/25	61/64	63/64

Table 7: 2-dimensional Cauchy target: errors of the marginal distributions of x_1 and x_2 for the seven pre-trained models before and after fine-tuning, and the error-reduction factors (before / after). Non-improvements are in gray; improvements are in bold. The last row reports the geometric mean of the factors over the seven models. t-EDM with $\nu = 3$.

	$\mathcal{E}_{L^1}$						$\mathcal{E}_{\text{tail}}$					
	x_1			x_2			x_1			x_2		
pre-trained model	before	after	factor	before	after	factor	before	after	factor	before	after	factor
Lip-KL-GPA	3.91	0.873	**4.5**×	5.11	0.858	**6.0**×	3.37	0.0578	**58.3**×	3.34	0.0788	**42.4**×
TTF	6.66	1.98	**3.4**×	11.4	5.94	**1.9**×	0.159	0.0786	**2.0**×	0.114	0.142	0.8×
Tail-GAN	25.0	1.33	**18.8**×	37.7	1.49	**25.3**×	4.63	0.129	**35.9**×	4.17	0.055	**75.8**×
t-EDM	3.34	1.29	**2.6**×	4.12	0.822	**5.0**×	0.19	0.0502	**3.8**×	0.24	0.049	**4.9**×
SHD	11.5	21.4	0.5×	275	45.5	**6.0**×	0.889	1	0.9×	2.89	0.987	**2.9**×
DLPM	482	10.2	**47.3**×	273	12.0	**22.8**×	0.644	0.42	**1.5**×	0.679	0.286	**2.4**×
mTAF	32.5	7.92	**4.1**×	54.4	10.0	**5.4**×	1.73	0.29	**6.0**×	1.79	0.355	**5.0**×
geometric mean			**5.0**×			**7.3**×			**5.4**×			**6.4**×

complete daily record over 2005–2024. Small catchments respond fast to rainfall, which is why their daily discharge has the heaviest tails in the basin. Tables 11 and 12 list the global and the tail error of every gauge for the seven pre-trained models and the fine-tuned models.

Table 8: Neal's funnel: errors of the marginal distributions of x_1 and x_2 for the seven pre-trained models before and after fine-tuning, and the error-reduction factors (before / after). The single non-improvement is in gray; improvements are in bold. The last row reports the geometric mean of the factors over the seven models. t-EDM with $\nu = (4, 20)$.

	$\mathcal{E}_{L^1}$						$\mathcal{E}_{\text{tail}}$					
	x_1			x_2			x_1			x_2		
pre-trained model	before	after	factor	before	after	factor	before	after	factor	before	after	factor
Lip-KL-GPA	0.0164	0.0108	**1.5**×	0.108	0.0228	**4.7**×	0.0394	0.0403	0.98×	0.0773	0.0188	**4.1**×
TTF	0.106	0.037	**2.9**×	0.47	0.0269	**17.5**×	0.37	0.0763	**4.8**×	0.218	0.0133	**16.4**×
Tail-GAN	0.352	0.0143	**24.6**×	1.48	0.0228	**64.9**×	1.02	0.0375	**27.2**×	3.73	0.0315	**118.4**×
t-EDM	0.135	0.0131	**10.3**×	0.102	0.021	**4.9**×	0.447	0.0196	**22.8**×	0.0768	0.0125	**6.1**×
SHD	0.158	0.0112	**14.1**×	1.71	0.0251	**68.1**×	0.884	0.0317	**27.9**×	0.576	0.0195	**29.5**×
DLPM	5.69	0.0146	**389.7**×	5.23	0.0303	**172.6**×	0.203	0.0312	**6.5**×	0.232	0.0216	**10.7**×
mTAF	0.0565	0.0093	**6.1**×	0.513	0.0206	**24.9**×	0.223	0.0427	**5.2**×	0.421	0.0193	**21.8**×
geometric mean			**12.1**×			**25.8**×			**8.3**×			**16.5**×

Table 11: Streamflow: $\mathcal{E}_{L^1}$ of each gauge (USGS site number, sorted by drainage area) with its Hill tail index $\hat{\alpha}$, for the seven pre-trained models (pre.) and the fine-tuned models (f.t.).

		Lip-KL-GPA		TTF		Tail-GAN		t-EDM		SHD		DLPM		mTAF	
gauge	$\hat{\alpha}$	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.
03284525	1.81	0.306	0.21	0.568	0.155	1.19	0.362	0.434	0.134	16.8	0.31	479	0.577	0.255	0.244
03357350	1.22	0.316	0.164	0.304	0.247	1.19	0.422	0.23	0.157	130	0.466	425	0.654	0.273	0.162
03301900	1.50	0.364	0.104	2.75	0.0955	1.12	0.201	0.357	0.116	44.2	0.3	298	0.367	0.298	0.148
03260100	1.42	0.508	0.126	0.48	0.233	1.8	0.417	0.301	0.21	37.2	0.198	506	0.779	0.49	0.213
03337000	1.82	0.124	0.133	0.398	0.112	0.439	0.27	0.106	0.0443	5.5	0.173	217	0.575	0.0499	0.0564
03337100	1.86	0.15	0.167	0.644	0.154	0.539	0.336	0.141	0.0707	2.12	0.225	264	0.665	0.068	0.0796
03298135	1.58	0.436	0.0633	5.05	0.214	1.17	0.262	0.408	0.119	81.5	0.252	539	0.523	0.322	0.112
03049800	1.84	0.156	0.109	0.103	0.0927	0.32	0.293	0.153	0.0715	2.36	0.136	274	0.416	0.161	0.123
03298300	1.47	0.338	0.0977	1143	0.143	1.45	0.213	0.28	0.118	125	0.215	285	0.411	0.299	0.133
03353120	1.38	0.252	0.072	0.203	0.121	1.51	0.37	0.174	0.151	44.9	0.216	362	0.62	0.246	0.0798
03026500	2.31	0.147	0.109	0.131	0.0665	0.462	0.179	0.144	0.0608	4.87	0.137	256	0.338	0.158	0.0877
03298550	1.39	0.296	0.172	1640	0.254	1.27	0.378	0.216	0.136	136	0.41	398	0.539	0.315	0.12
03274650	1.43	0.237	0.118	0.241	0.16	1	0.301	0.128	0.111	11.0	0.0979	229	0.435	0.281	0.101
03298150	1.45	0.418	0.102	3.31	0.155	1.31	0.185	0.372	0.0742	13.6	0.142	340	0.381	0.35	0.113
03431300	1.67	0.245	0.162	0.593	0.226	1.17	0.328	0.285	0.0982	51.3	0.258	428	0.564	0.397	0.2
03611260	1.16	0.352	0.222	118	0.235	1.75	0.742	0.285	0.197	252	0.724	612	1.12	0.55	0.298
03353620	1.40	0.313	0.0974	7.55	0.156	1.18	0.238	0.189	0.119	35.6	0.213	291	0.451	0.257	0.102
03351310	1.53	0.189	0.112	0.534	0.0789	0.791	0.199	0.173	0.101	7.86	0.117	245	0.366	0.204	0.123
03293000	1.48	0.326	0.0736	1.58	0.105	1.02	0.183	0.324	0.0747	16.4	0.168	275	0.358	0.228	0.11
03302680	1.53	0.292	0.263	0.721	0.149	1.02	0.489	0.224	0.157	69.6	0.374	302	0.673	0.286	0.154
03289000	2.04	0.17	0.0837	0.242	0.0881	0.311	0.146	0.292	0.0549	43.4	0.21	301	0.326	0.126	0.0757
03262001	1.34	0.342	0.0837	0.273	0.1	1.18	0.195	0.187	0.0895	62.7	0.191	284	0.4	0.449	0.161
03431700	1.66	0.182	0.102	0.326	0.114	0.923	0.212	0.238	0.0825	12.5	0.117	357	0.429	0.367	0.176
03353600	1.78	0.21	0.0667	0.148	0.0567	0.655	0.186	0.155	0.115	67.5	0.185	263	0.348	0.211	0.0908
03140000	1.35	0.203	0.155	0.145	0.119	0.84	0.317	0.11	0.0722	2.43	0.297	217	0.48	0.192	0.166
03291780	1.37	0.332	0.162	0.354	0.143	1.25	0.298	0.163	0.123	130	0.323	346	0.491	0.382	0.151
03289200	2.30	0.304	0.148	0.709	0.231	0.794	0.222	0.387	0.0838	16.8	0.313	369	0.373	0.205	0.106
03254550	1.53	0.288	0.109	0.224	0.0946	0.96	0.238	0.229	0.0919	42.4	0.218	278	0.408	0.296	0.117
03213500	1.54	0.163	0.156	1.22	0.129	0.633	0.286	0.109	0.0832	27.7	0.189	259	0.435	0.208	0.111
03339280	1.55	0.173	0.115	0.148	0.126	0.519	0.217	0.12	0.0688	3.46	0.0611	196	0.376	0.209	0.1
03230450	1.79	0.178	0.122	0.253	0.137	0.517	0.197	0.156	0.0671	11.1	0.117	225	0.352	0.226	0.0868
03266560	1.80	0.165	0.12	0.319	0.0348	0.373	0.152	0.126	0.0475	8.5	0.14	167	0.273	0.112	0.0769
03277075	1.49	0.342	0.0673	0.353	0.119	1.13	0.151	0.248	0.079	5.4	0.185	254	0.355	0.32	0.0951
03223425	1.70	0.183	0.149	1.87	0.167	0.727	0.293	0.148	0.0957	6.91	0.37	269	0.447	0.219	0.108
03011800	2.64	0.178	0.12	0.346	0.114	0.481	0.174	0.188	0.0548	74.7	0.21	253	0.31	0.199	0.113
03238745	1.28	0.325	0.144	0.418	0.18	1.65	0.311	0.163	0.0971	112	0.324	284	0.543	0.372	0.151
03260706	1.83	0.15	0.121	7.91	0.101	0.205	0.146	0.131	0.0534	6.88	0.0984	168	0.291	0.134	0.0678
03384450	1.05	0.409	0.215	2.61	0.331	1.64	0.523	0.268	0.212	84.7	0.46	438	0.81	0.386	0.237
03177710	2.22	0.141	0.153	0.145	0.0761	0.263	0.189	0.168	0.0512	2.84	0.0797	227	0.326	0.0897	0.0706
03076600	2.13	0.177	0.12	0.171	0.0929	0.335	0.173	0.296	0.0697	2.92	0.198	282	0.376	0.203	0.0978
03330241	2.92	0.154	0.152	0.105	0.089	0.22	0.203	0.0737	0.033	5.9	0.18	148	0.282	0.0506	0.0341
03350700	1.66	0.149	0.115	0.0931	0.0587	0.628	0.203	0.138	0.0745	52.1	0.197	192	0.322	0.165	0.0776
03431599	1.42	0.188	0.157	0.24	0.207	0.899	0.331	0.232	0.117	16.6	0.151	364	0.583	0.253	0.145
03106300	3.34	0.106	0.111	0.154	0.0625	0.42	0.156	0.129	0.0509	3.94	0.179	200	0.266	0.0386	0.0405
03185000	1.80	0.149	0.133	0.208	0.0702	0.271	0.208	0.134	0.0556	30.9	0.173	269	0.332	0.122	0.1
03244936	1.31	0.364	0.185	0.317	0.194	1.59	0.365	0.232	0.13	242	0.374	367	0.563	0.431	0.181
03210000	1.75	0.13	0.154	1.08	0.129	0.587	0.242	0.0999	0.0682	10.6	0.11	224	0.383	0.181	0.117
03042280	2.24	0.157	0.129	1.01	0.108	0.425	0.214	0.196	0.0726	53.4	0.21	333	0.393	0.133	0.109
03067510	2.03	0.285	0.194	0.381	0.211	0.619	0.352	0.481	0.114	29.0	0.343	519	0.593	0.296	0.193
03280700	1.73	0.139	0.132	0.334	0.12	0.625	0.257	0.113	0.0771	5.59	0.283	266	0.386	0.188	0.0973
03062500	2.75	0.0959	0.0788	0.165	0.0791	0.348	0.154	0.164	0.0679	69.1	0.142	266	0.273	0.143	0.0493
03241500	1.74	0.146	0.105	0.14	0.0545	0.39	0.171	0.165	0.056	4.98	0.135	194	0.311	0.196	0.0981
03302000	1.63	0.284	0.0964	0.277	0.0591	0.829	0.202	0.196	0.0935	5.13	0.0668	314	0.315	0.236	0.0803
03078000	2.21	0.177	0.121	0.201	0.117	0.402	0.161	0.247	0.0521	132	0.259	268	0.351	0.242	0.111
03228750	1.43	0.223	0.123	8.51	0.178	1.01	0.363	0.18	0.117	42.0	0.197	266	0.535	0.226	0.132
03302220	1.68	0.211	0.0839	0.273	0.116	0.817	0.186	0.193	0.0744	15.3	0.127	260	0.308	0.22	0.0627
03277300	1.97	0.101	0.0979	0.284	0.086	0.207	0.164	0.114	0.0422	8.89	0.139	207	0.321	0.15	0.0807
03208950	1.83	0.102	0.105	0.35	0.0807	0.451	0.158	0.0991	0.0454	2.42	0.164	191	0.327	0.177	0.0995
03271000	1.32	0.247	0.114	0.162	0.0675	0.783	0.199	0.169	0.0757	6.07	0.0599	213	0.34	0.258	0.0819
03272700	1.44	0.237	0.108	0.133	0.0754	0.712	0.216	0.177	0.13	50.6	0.162	207	0.348	0.296	0.104
03141870	1.73	0.13	0.111	0.238	0.145	0.573	0.252	0.193	0.0937	5.57	0.187	192	0.363	0.147	0.0874
03074500	2.24	0.133	0.1	0.215	0.0831	0.324	0.146	0.139	0.0845	5.49	0.212	200	0.302	0.13	0.0625
03149500	1.51	0.136	0.124	0.333	0.146	0.541	0.211	0.164	0.0829	3.75	0.161	174	0.337	0.202	0.096
03361850	1.75	0.186	0.088	0.497	0.114	0.902	0.211	0.17	0.0657	4.17	0.173	289	0.34	0.177	0.0652

Table 12: Streamflow: $\mathcal{E}_{\text{tail}}$ of each gauge (USGS site number, sorted by drainage area) with its Hill tail index $\hat{\alpha}$, for the seven pre-trained models (pre.) and the fine-tuned models (f.t.).

		Lip-KL-GPA		TTF		Tail-GAN		t-EDM		SHD		DLPM		mTAF	
gauge	$\hat{\alpha}$	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.	pre.	f.t.
03284525	1.81	2.08	0.248	0.372	0.142	1.5	0.286	0.552	0.067	3.04	0.266	3.95	0.325	0.323	0.18
03357350	1.22	1.41	0.231	0.626	0.27	1.76	0.296	0.264	0.0961	3.79	0.426	3.95	0.397	0.427	0.143
03301900	1.50	2.31	0.107	0.167	0.0385	1.28	0.227	1.25	0.132	3.17	0.304	3.93	0.258	1.16	0.152
03260100	1.42	1.81	0.0875	0.475	0.131	1.25	0.118	0.91	0.131	2.6	0.158	3.94	0.194	1.28	0.221
03337000	1.82	0.138	0.122	0.29	0.168	0.364	0.162	0.205	0.0638	1.66	0.213	3.95	1.22	0.146	0.0906
03337100	1.86	0.242	0.148	0.308	0.208	0.437	0.164	0.225	0.126	0.606	0.101	3.95	1.26	0.17	0.128
03298135	1.58	2.3	0.0582	0.343	0.176	0.392	0.126	1.32	0.104	3.3	0.112	3.94	0.15	1.05	0.058
03049800	1.84	1.36	0.229	0.16	0.121	0.365	0.903	0.198	0.174	1.06	0.218	3.95	0.345	0.397	1.01
03298300	1.47	3.08	0.0832	0.299	0.0998	1.46	0.109	0.362	0.0831	3.81	0.137	3.93	0.169	1.14	0.118
03353120	1.38	1.45	0.102	0.212	0.0926	1.36	0.123	0.111	0.11	3.25	0.212	3.95	0.235	0.452	0.113
03026500	2.31	1.13	0.252	0.194	0.0702	0.769	0.924	0.248	0.102	2.28	0.0942	3.95	1	1.1	0.245
03298550	1.39	2.44	0.127	0.269	0.267	0.687	0.154	0.332	0.0861	3.84	0.203	3.95	0.203	0.511	0.148
03274650	1.43	2.36	0.108	0.529	0.126	1.33	0.12	0.135	0.0804	2.28	0.0639	3.93	0.266	1.17	0.172
03298150	1.45	2.45	0.0472	0.285	0.129	1.24	0.0821	0.553	0.0655	2.37	0.116	3.93	0.0963	1.24	0.102
03431300	1.67	1.18	0.132	0.179	0.232	1.43	0.189	0.245	0.107	3.71	0.279	3.95	0.327	0.523	0.108
03611260	1.16	1.42	0.135	0.6	0.117	1.44	0.353	0.414	0.113	3.91	0.293	3.95	0.468	1.46	0.361
03353620	1.40	2.26	0.145	0.0956	0.162	1.49	0.16	0.127	0.147	3.43	0.262	3.93	0.344	1.11	0.122
03351310	1.53	2.06	0.205	0.168	0.0964	1.26	0.271	0.157	0.135	2.83	0.141	3.94	1.16	1.16	0.188
03293000	1.48	3.17	0.0778	0.179	0.109	1.37	0.105	1.25	0.0675	3.35	0.127	3.94	0.194	1.08	0.114
03302680	1.53	2.4	0.314	0.165	0.0877	1.43	0.4	0.151	0.0569	3.26	0.47	3.94	0.437	0.499	0.239
03289000	2.04	2.18	0.143	0.319	0.11	0.474	0.152	0.464	0.0594	3.17	0.304	3.95	0.33	0.26	0.105
03262001	1.34	3.34	0.103	0.459	0.0682	1.48	0.0891	0.973	0.0972	3.57	0.12	3.93	0.134	2.22	0.234
03431700	1.66	1.13	0.122	0.227	0.115	1.42	0.134	0.336	0.0797	3.17	0.0946	3.95	0.322	0.584	0.132
03353600	1.78	2.2	0.0888	0.157	0.0746	0.481	0.0918	0.115	0.107	3.43	0.249	3.94	0.195	1.18	0.139
03140000	1.35	1.44	0.258	0.209	0.167	1.54	0.332	0.115	0.0529	1.05	0.337	3.93	1.23	0.426	0.263
03291780	1.37	2.67	0.127	0.372	0.0998	1.56	0.156	0.0959	0.0469	3.76	0.166	3.94	0.271	1.33	0.148
03289200	2.30	3.11	0.121	0.426	0.333	0.678	0.0943	0.421	0.058	2.59	0.196	3.94	0.141	0.297	0.0971
03254550	1.53	3.16	0.128	0.361	0.112	1.36	0.193	0.25	0.0773	3.26	0.267	3.94	1.04	0.442	0.149
03213500	1.54	1.14	0.205	0.188	0.167	1.36	0.204	0.247	0.117	3.08	0.354	3.95	0.358	1.24	0.229
03339280	1.55	2.28	0.0837	0.107	0.202	1.3	0.121	0.0345	0.0448	1.38	0.0649	3.93	0.239	1.15	0.145
03230450	1.79	2.18	0.127	0.193	0.255	0.459	0.155	0.215	0.0634	2.62	0.136	3.94	0.956	1.19	0.945
03266560	1.80	2.2	0.225	0.395	0.074	0.785	0.208	1.01	0.121	2.75	0.27	3.95	0.546	0.409	0.195
03277075	1.49	3.36	0.0335	0.504	0.113	0.46	0.0581	0.243	0.0683	1.48	0.0582	3.93	0.16	2.09	0.118
03223425	1.70	2.26	0.133	0.204	0.233	0.589	0.145	0.152	0.1	2.47	0.175	3.94	0.265	0.412	0.178
03011800	2.64	1.14	0.222	0.266	0.141	0.854	0.951	0.179	0.0679	3.79	0.317	3.95	1.01	1.16	0.232
03238745	1.28	4.08	0.145	0.523	0.159	1.21	0.167	0.148	0.0642	3.63	0.183	3.93	0.259	1.3	0.257
03260706	1.83	2.14	0.202	0.0898	0.22	0.507	0.198	0.274	0.084	2.36	0.196	3.94	0.43	0.36	0.176
03384450	1.05	2.3	0.125	0.582	0.235	1.66	0.248	0.273	0.129	3.6	0.306	3.94	0.346	1.29	0.29
03177710	2.22	0.421	0.188	0.0902	0.134	0.688	0.253	0.269	0.147	1.84	0.127	3.95	0.347	1.08	0.22
03076600	2.13	1.26	0.185	0.15	0.152	0.537	0.197	0.237	0.0841	1.06	0.199	3.94	1.01	0.353	0.181
03330241	2.92	2.26	1.43	0.318	0.333	0.939	3.84	0.223	0.0946	2.56	0.39	3.96	0.349	1.11	1.03
03350700	1.66	2.97	0.174	0.082	0.0732	0.6	0.178	0.134	0.112	3.36	0.293	3.93	0.288	0.3	0.106
03431599	1.42	1.13	0.147	0.104	0.194	1.45	0.18	0.294	0.101	3.27	0.137	3.95	0.403	0.363	0.122
03106300	3.34	1.46	0.47	0.339	0.213	1.24	0.707	0.372	0.171	3.03	0.527	3.96	0.384	0.249	0.21
03185000	1.80	1.15	0.237	0.22	0.0941	0.592	0.204	0.196	0.137	3.17	0.392	3.96	0.335	1.1	0.25
03244936	1.31	3.23	0.193	0.242	0.166	1.47	0.177	0.336	0.0789	3.95	0.174	3.94	0.987	1.43	0.281
03210000	1.75	1.15	0.203	0.139	0.251	0.71	0.176	0.219	0.116	2.81	0.137	3.95	0.386	1.22	0.257
03042280	2.24	1.24	0.232	0.261	0.184	0.843	0.961	0.162	0.0726	3.35	0.435	3.96	1.11	1.08	0.227
03067510	2.03	1.26	0.145	0.188	0.283	0.651	1.03	1.12	0.0986	3.02	0.316	3.95	1.17	0.425	0.241
03280700	1.73	1.18	0.168	0.268	0.188	0.626	0.183	0.0977	0.079	2.37	0.211	3.94	0.275	0.42	0.152
03062500	2.75	1.35	0.12	0.229	0.19	0.702	0.192	0.166	0.0805	3.53	0.133	3.95	0.169	0.291	0.119
03241500	1.74	1.53	0.117	0.151	0.0984	0.527	0.102	0.239	0.0617	2.43	0.0938	3.95	1.01	0.465	0.182
03302000	1.63	3.29	0.127	0.13	0.0516	1.22	0.0984	1.06	0.0809	1.66	0.0517	3.94	0.106	1.1	0.123
03078000	2.21	1.27	0.102	0.165	0.188	0.603	0.164	0.212	0.0523	3.93	0.211	3.94	0.282	0.42	0.129
03228750	1.43	2.31	0.123	0.286	0.194	1.28	0.195	0.317	0.0842	3.43	0.272	3.94	1.05	0.476	0.243
03302220	1.68	3.26	0.0788	0.185	0.173	1.36	0.0845	1.14	0.0581	2.92	0.149	3.94	0.119	1.19	0.0971
03277300	1.97	1.07	0.111	0.183	0.203	1.26	0.107	0.233	0.0489	3.01	0.203	3.94	0.358	0.446	0.174
03208950	1.83	1.04	0.173	0.156	0.184	1.5	0.108	0.169	0.0754	1.62	0.215	3.94	0.462	0.573	0.217
03271000	1.32	3.18	0.149	0.345	0.0845	1.34	0.176	0.121	0.118	2.3	0.0558	3.93	0.294	1.25	0.175
03272700	1.44	2.46	0.0963	0.166	0.0925	1.19	0.107	0.22	0.168	3.4	0.211	3.93	0.224	1.23	0.149
03141870	1.73	1.38	0.138	0.181	0.283	0.476	0.153	0.413	0.0766	1.74	0.297	3.94	0.286	0.396	0.165
03074500	2.24	1.98	0.932	0.12	0.138	0.737	0.887	0.133	0.185	2.77	0.98	3.95	0.453	1.06	0.188
03149500	1.51	2.14	0.258	0.108	0.224	1.3	0.18	0.355	0.089	1.99	0.188	3.93	1.03	1.18	0.214
03361850	1.75	2.36	0.154	0.162	0.171	0.501	0.202	0.0989	0.0628	1.5	0.23	3.94	0.219	1.03	0.122

D CVaR-PENALIZED WASSERSTEIN GRADIENT FLOWS WITH CLARKE'S GENERALIZED SUBGRADIENTS

In this Appendix, we analyze the CVaR-penalized Wasserstein gradient flows defined using CVaR (as opposed to smoothed CVaR). All the relevant proofs, and the statements of auxiliary lemmas accompanied by their proofs, are presented in Appendix D.5. We define the loss functional $\mathcal{F}$

$$\mathcal{F}(Q; P^{\mathrm{tar}}) = D_{\mathrm{KL}}^{L}(Q \| P^{\mathrm{tar}}) + \lambda \left(\Delta C(Q; P^{\mathrm{tar}})\right)^2, \tag{49}$$

where $\Delta C(Q; P^{\mathrm{tar}}) := \mathrm{CVaR}_{\alpha}^{P^{\mathrm{tar}},g} - \mathrm{CVaR}_{\alpha}^{Q,g}$ denotes the CVaR discrepancy of $g(X)$ for $X \sim Q$ and $X \sim P^{\mathrm{tar}}$, for a given risk function $g : \mathbb{R}^d \to [0, \infty)$.

Definition D.1. *Let $g : \mathbb{R}^d \to [0, \infty)$ be a risk function, and $Q \in \mathcal{Q}^g$. Given a parameter $0 < \alpha < 1$, the upper and lower α-quantiles under Q are given by*

$$\overline{\mathrm{VaR}}_{\alpha}^{Q,g} := \inf\{c : \Psi^{Q,g}(c) > \alpha\},$$
$$\mathrm{VaR}_{\alpha}^{Q,g} := \inf_{c \in \mathbb{R}}\{c : \Psi^{Q,g}(c) \geq \alpha\},$$

where $\Psi^{Q,g}$ denotes the cumulative distribution function of g under Q.

CVaR is defined via the Rockafellar–Uryasev formulation (Rockafellar & Uryasev, 2002) (see also Equation 4)

$$\mathrm{CVaR}_{\alpha}^{Q,g} = \inf_{y \in \mathbb{R}} F_{\alpha}^{Q,g}(y) := \inf_{y \in \mathbb{R}} \left\{ y + \frac{1}{1-\alpha} \mathbb{E}_Q[(g(\cdot) - y)^+] \right\}. \tag{50}$$

Then the set of minimizers of Equation 50, denoted by $T(Q)$, is $[\mathrm{VaR}_{\alpha}^{Q,g}, \overline{\mathrm{VaR}}_{\alpha}^{Q,g}]$ (see Rockafellar & Uryasev (2002, Theorem 10) for a proof).

Under Assumption B.1, the loss functional $\mathcal{F}$ is well defined, and the argument of Proposition B.3 shows that it is a divergence, so that P^{tar} is the unique global minimizer of the optimization problem $\inf_{Q \in \mathcal{Q}^g} \mathcal{F}(Q; P^{\mathrm{tar}})$. Its Wasserstein gradient flow, however, requires a variational derivative, which does not exist in general: $\mathrm{CVaR}_{\alpha}^{Q,g}$ is Gateaux differentiable at Q if and only if $\mathrm{VaR}_{\alpha}^{Q,g} = \overline{\mathrm{VaR}}_{\alpha}^{Q,g}$ (Theorem D.5(4)), and the same holds for $\mathcal{F}$ whenever $\Delta C(Q; P^{\mathrm{tar}}) \neq 0$ (Theorem D.7(3)). This condition fails whenever $\Psi^{Q,g} = \alpha$ on a nondegenerate interval, and, for an empirical measure, whose CDF $\Psi^{Q,g}$ is a step function, such levels α always exist. We therefore work with variational subgradients, defined via Clarke's generalized gradients (Clarke, 1990).

D.1 CLARKE'S GENERALIZED SUBGRADIENTS

Let $Q \mapsto J(Q)$ be a locally Lipschitz continuous (with respect to the Wasserstein-1 metric $\mathcal{W}_1$ on $\mathcal{P}(\mathbb{R}^d)$) real-valued functional, defined on $\mathcal{P}(\mathbb{R}^d)$. The Clarke generalized directional derivative (Clarke, 1990) of J at Q for an admissible perturbation ρ (Definition B.2) is defined by

$$J^{\circ}(Q; \rho) := \overline{\lim_{Q' \to Q,\ t \downarrow 0}} \frac{J(Q' + t\rho) - J(Q')}{t}. \tag{51}$$

The *Clarke subdifferential* of J at Q is defined as

$$\partial J(Q) := \left\{ \xi :\ J^{\circ}(Q; \rho) \geq \int \xi \, d\rho \ \text{ for all admissible perturbations } \rho \right\}, \tag{52}$$

and its elements are referred to as the first variational subgradients of J at Q.

Convex and concave cases When J is *convex*, its Clarke subdifferential coincides with the classical *subdifferential* of convex analysis

$$\partial J(Q) := \left\{ \xi :\ J(Q') \geq J(Q) + \int \xi \, d(Q' - Q) \ \ \forall Q' \in \mathcal{P}(\mathbb{R}^d) \right\}. \tag{53}$$

When J is *concave*, the Clarke subdifferential instead coincides with the *superdifferential*

$$\partial J(Q) := \left\{ \xi :\ J(Q') \leq J(Q) + \int \xi \, d(Q' - Q) \ \ \forall Q' \in \mathcal{P}(\mathbb{R}^d) \right\}, \tag{54}$$

and its elements are correspondingly called *supergradients* rather than subgradients.

The risk measure $\mathrm{CVaR}_\alpha^{Q,g}$ is an infimum of a family of functions affine in Q, by the Rockafellar–Uryasev formulation Equation 50. Thus $\mathrm{CVaR}_\alpha^{\cdot,g}$ is concave in Q. However, in this paper we use the single term *variational subgradients* throughout to refer to the elements of ∂J for any given functional J, regardless of whether J is convex, concave, or neither.

Local Lipschitz continuity with respect to the Wasserstein-1 metric Clarke's generalized directional derivative is finite when J is locally Lipschitz continuous with respect to a metric on $\mathcal{P}(\mathbb{R}^d)$ (Lemma D.2). Since the perturbation $Q^\epsilon = Q + \epsilon\rho$ displaces probability mass, we use the Wasserstein-1 metric, which, unlike metrics insensitive to spatial displacement such as the total variation norm, measures the cost of this displacement. Wasserstein gradient flows are classically formulated in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ (Jordan et al., 1998; Otto, 2001); since $\mathcal{W}_1 \le \mathcal{W}_2$, local Lipschitz continuity with respect to $\mathcal{W}_1$ implies local Lipschitz continuity with respect to $\mathcal{W}_2$.

Lemma D.2. *Let $J(Q)$ be locally Lipschitz continuous with respect to the Wasserstein-1 metric on $\mathcal{P}(\mathbb{R}^d)$. Then for any admissible perturbation ρ, the Clarke generalized directional derivative $J^\circ(Q;\rho)$ exists and is finite.*

Throughout this appendix, the risk function g is Lipschitz continuous. This ensures that $Q \mapsto \mathcal{F}(Q; P^{\mathrm{tar}})$ is locally Lipschitz continuous with respect to $\mathcal{W}_1$, so that its variational subgradients are well defined.

Proposition D.3 (Local Lipschitz continuity)**.** *Assume that g is $\mathrm{Lip}(g)$-Lipschitz continuous and Assumption B.1 holds. Then the functional $Q \mapsto \mathcal{F}(Q; P^{\mathrm{tar}})$ is locally Lipschitz continuous on $\mathcal{Q}^g$ with respect to $\mathcal{W}_1$.*

The variational subgradients of $\mathrm{CVaR}_\alpha^{Q,g}$ and $\mathcal{F}(Q; P^{\mathrm{tar}})$ are identified through the following lemma.

Lemma D.4. *Let $J(Q)$ be locally Lipschitz continuous with respect to the Wasserstein-1 metric on $\mathcal{P}(\mathbb{R}^d)$. Define the right/left one-sided directional derivatives of J as:*

$$D_\rho^\pm J := \lim_{\epsilon \to 0^\pm} \frac{1}{\epsilon}\big(J(Q^\epsilon) - J(Q)\big). \tag{55}$$

Assume that for every admissible perturbation ρ at Q, the one-sided directional derivatives $D_\rho^\pm J$ exist and $\int \xi \, d\rho$ lies in the closed interval with endpoints $D_\rho^- J$ and $D_\rho^+ J$. Then ξ is a variational subgradient of J at Q.

Variational subgradients of $\mathbf{CVaR}_{\boldsymbol{\alpha}}^{\boldsymbol{Q,g}}$ Since g is non-negative and Lipschitz continuous, $0 \le g(x) \le g(0) + \mathrm{Lip}(g)\|x\|$, so that the condition $\int g \, d|\rho| < \infty$ in Definition B.2 follows from $\int \|x\| \, d|\rho| < \infty$:

$$\int g \, d|\rho| \;\le\; g(0) \int d|\rho| \;+\; \mathrm{Lip}(g) \int \|x\| \, d|\rho| \;<\; \infty. \tag{56}$$

In particular, by Equation 13, $\mathrm{CVaR}_\alpha^{Q^\epsilon,g}$ is well defined for every right- (left-) admissible ρ and all small $\epsilon > 0$ ($\epsilon < 0$).

Theorem D.5 (Directional derivatives and variational subgradients of CVaR)**.** *Assume that g is Lipschitz continuous and Q has finite first moment (hence $Q \in \mathcal{Q}^g$). For $y \in \mathbb{R}$, define $h^y(x) := \frac{(g(x)-y)^+}{1-\alpha}$. Then we have the following:*

1. (Right derivative) *For every right-admissible perturbation ρ, the right directional derivative exists and is given by*

$$D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g} = \lim_{\epsilon \to 0^+} \frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} = \min_{y \in T(Q)} \int h^y \, d\rho. \tag{57}$$

2. (Left derivative) *For every left-admissible perturbation ρ, the left directional derivative exists and is given by*

$$D_\rho^- \mathrm{CVaR}_\alpha^{Q,g} = \lim_{\epsilon \to 0^-} \frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} = \max_{y \in T(Q)} \int h^y \, d\rho. \tag{58}$$

3. (Variational subgradients) *Let $y \in T(Q)$. Then, for any admissible ρ, $\int h^y \, d\rho$ lies in $[D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g}, D_\rho^- \mathrm{CVaR}_\alpha^{Q,g}]$. Consequently, for each $y \in T(Q)$, h^y is a variational subgradient of $\mathrm{CVaR}_\alpha^{Q,g}$.*

4. (Two-sided derivative) *The two-sided (Gateaux) derivative for an admissible ρ exists if and only if $y \mapsto \int h^y \, d\rho$ is constant on $T(Q)$; this holds for every admissible ρ if and only if $T(Q)$ is a singleton set, i.e. $\mathrm{VaR}_\alpha^{Q,g} = \overline{\mathrm{VaR}}_\alpha^{Q,g}$, in which case $\mathrm{CVaR}_\alpha^{Q,g}$ is Gateaux differentiable at Q with the first variation*

$$\frac{\delta \mathrm{CVaR}_\alpha^{Q,g}}{\delta Q} = h^{\mathrm{VaR}_\alpha^{Q,g}} = \frac{(g - \mathrm{VaR}_\alpha^{Q,g})^+}{1-\alpha}.$$

Remark D.6. *For the radial risk function $g(x) = \|x\|$, the hypotheses of Theorem D.5 reduce to $Q \in \mathcal{Q}^g$, which is exactly what is needed for $\mathrm{CVaR}_\alpha^{Q,g}$ to be well defined; no hypothesis on the set $T(Q)$ is imposed. Part (4) recovers the formula $\frac{\delta \mathrm{CVaR}_\alpha^{Q,g}}{\delta Q} = \frac{(g-\mathrm{VaR}_\alpha^{Q,g})^+}{1-\alpha}$ employed in Santi et al. (2025); Wang et al. (2026), where it is derived under a density assumption on the law of $g(X)$, $X \sim Q$, that makes $T(Q)$ a singleton. This assumption fails for empirical measures, for which the variational subgradients of part (3) are required.*

We now consider the variational subgradients of the loss functional $\mathcal{F}$ itself.

Theorem D.7 (Directional derivatives and variational subgradients of $\mathcal{F}$)**.** *Assume that g is Lipschitz continuous and Assumption B.1 holds. Let ϕ^* denote the maximizer of Equation 3. For each $y \in T(Q)$, define the potential function*

$$\Phi_Q^y := \phi^* - 2\lambda \, \Delta C(Q; P^{\mathrm{tar}}) \, h^y = \phi^* - \frac{2\lambda}{1-\alpha} \, \Delta C(Q; P^{\mathrm{tar}}) \, (g-y)^+. \tag{59}$$

Then the following hold.

1. *For every right-admissible perturbation ρ, the right directional derivative exists and is given by*

$$D_\rho^+ \mathcal{F}(Q; P^{\mathrm{tar}}) = \int \phi^* \, d\rho - 2\lambda \, \Delta C(Q; P^{\mathrm{tar}}) \min_{y \in T(Q)} \int h^y d\rho. \tag{60}$$

For every left-admissible perturbation ρ, the left directional derivative exists and is given by

$$D_\rho^- \mathcal{F}(Q; P^{\mathrm{tar}}) = \int \phi^* \, d\rho - 2\lambda \, \Delta C(Q; P^{\mathrm{tar}}) \max_{y \in T(Q)} \int h^y d\rho. \tag{61}$$

2. *For any admissible ρ, the range of $\int \Phi_Q^y d\rho$ for $y \in T(Q)$ is the compact interval between $D_\rho^- \mathcal{F}(Q; P^{\mathrm{tar}})$ and $D_\rho^+ \mathcal{F}(Q; P^{\mathrm{tar}})$. Consequently, the family of potential functions $\{\Phi_Q^y; y \in T(Q)\}$ are subgradients of $\mathcal{F}(Q; P^{\mathrm{tar}})$.*

3. *Let ρ be admissible. The two-sided (Gateaux) derivative of $\mathcal{F}$ for the perturbation ρ exists if and only if $\Delta C(Q; P^{\mathrm{tar}}) = 0$ or $y \mapsto \int h^y \, d\rho$ is constant on $T(Q)$. The two-sided (Gateaux) derivative of $\mathcal{F}$ exists for every admissible perturbation if and only if $\Delta C(Q; P^{\mathrm{tar}}) = 0$ or $T(Q)$ is a singleton set; in that case, $\mathcal{F}(\cdot; P^{\mathrm{tar}})$ is Gateaux differentiable at Q with first variational derivative*

$$\frac{\delta \mathcal{F}(Q; P^{\mathrm{tar}})}{\delta Q} = \Phi_Q^{\mathrm{VaR}_\alpha^{Q,g}}.$$

D.2 Velocity field of the CVaR-penalized Wasserstein gradient flow

By Theorem D.7(3), the variational derivative of $\mathcal{F}(Q; P^{\mathrm{tar}})$, which defines the velocity field of its Wasserstein gradient flow, does not exist unless $T(Q)$ is a singleton or $\Delta C(Q; P^{\mathrm{tar}}) = 0$. We therefore define the velocity field through a variational subgradient. By Theorem D.7(2), every Φ_Q^y, $y \in T(Q)$, is a variational subgradient, and the choice of y is a genuine degree of freedom in the design of the flow.

Theorem D.8 (The velocity field of $\mathcal{F}$)**.** *Assume that g is Lipschitz continuous, $g \in C^1$ for $g(x) \neq 0$, and Assumption B.1 holds. For each $y \in T(Q)$, choosing the potential function Φ_Q^y induces the velocity field $v_Q^y = -\nabla_x \Phi_Q^y$, and we have*

$$v_Q^y(x) = \begin{cases} \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})} + \underbrace{\dfrac{2\lambda}{1-\alpha}\,\Delta C(Q;P^{\mathrm{tar}})\,\nabla_x g(x)}_{\text{from CVaR-penalization}}, & g(x) > y, \\ \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})}, & g(x) \le y. \end{cases} \tag{62}$$

If $T(Q)$ is a singleton set, the velocity field degenerates to $v_Q^{\mathrm{VaR}_\alpha^{Q,g}}$.

Corollary D.9 (Boundedness of the velocity field)**.** *Assume that g is Lipschitz continuous, $g \in C^1$ for $g(x) \neq 0$, and Assumption B.1 holds. For every $y \in T(Q)$, the velocity field v_Q^y in Equation 62 satisfies, for almost every $x \in \mathbb{R}^d$,*

$$\left\|v_Q^y(x)\right\| \;\le\; L + \frac{2\lambda}{1-\alpha}\,\mathrm{Lip}(g)\,\left|\Delta C(Q;P^{\mathrm{tar}})\right|. \tag{63}$$

Moreover, the CVaR discrepancy is controlled by the loss $\mathcal{F}(Q;P^{\mathrm{tar}})$, so we have

$$\left\|v_Q^y(x)\right\| \;\le\; L + \frac{2\,\mathrm{Lip}(g)}{1-\alpha}\,\sqrt{\lambda\,\mathcal{F}(Q;P^{\mathrm{tar}})}, \tag{64}$$

for almost every $x \in \mathbb{R}^d$. In particular, the velocity field is bounded, and the CVaR-penalization component vanishes as $\mathcal{F}(Q;P^{\mathrm{tar}}) \to 0$, i.e. as $Q \to P^{\mathrm{tar}}$.

Remark D.10 (Bounded but non-Lipschitz continuous)**.** *Although Corollary D.9 shows that the velocity field v_Q^y in Equation 62 is bounded, it is not Lipschitz continuous in x, for two reasons. First, v_Q^y has a jump of magnitude $\frac{2\lambda}{1-\alpha}\,|\Delta C(Q;P^{\mathrm{tar}})|\,\|\nabla_x g\|$ at the threshold set $\{x : g(x) = y\}$. Second, $\nabla_x \phi^*$ is bounded due to ϕ^* being Lipschitz continuous, but $\nabla_x \phi^*$ is not necessarily continuous.*

The bounded yet discontinuous velocity field v_Q^y in Equation 62 distinguishes the CVaR-penalized flow from generative models built on Lipschitz transport maps: the latter preserve the tail behavior of the light-tailed source distribution (Jaini et al., 2020), whereas a non-Lipschitz transport map is required to transport the pre-trained distribution P^{pre} toward the heavier-tailed target distribution P^{tar}. We do not claim here a rigorous quantitative change of the tail behavior; the resulting gain in tail accuracy is demonstrated empirically in Section 4.

Throughout this paper, we adopt the radial risk function $g(x) = \|x\|$, which is 1-Lipschitz continuous and C^1 on $\{x \neq 0\}$, so that Theorem D.8 applies and yields the following velocity field.

Corollary D.11. *Let Assumption B.1 hold and let $g(x) = \|x\|$ and $\mathbb{E}_Q[g] < \infty$. For each $y \in T(Q)$, choosing the potential function Φ_Q^y induces the velocity field $v_Q^y = -\nabla_x \Phi_Q^y$, and we have*

$$v_Q^y(x) = \begin{cases} \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})} + \underbrace{\dfrac{2\lambda}{1-\alpha}\,\Delta C(Q;P^{\mathrm{tar}})\,\dfrac{x}{\|x\|}}_{\text{from CVaR-penalization}}, & \|x\| > y, \\ \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})}, & \|x\| \le y. \end{cases} \tag{65}$$

D.3 The velocity contribution from CVaR-penalization

The CVaR-penalized Wasserstein gradient flow is given by

$$\partial_t Q_t + \nabla \cdot (Q_t v_{Q_t}^y) = 0, \quad Q_0 = P^{\mathrm{pre}}, \tag{66}$$

where the velocity field $v_{Q_t}^y$ is given by Equation 65. Its component induced by the CVaR-penalization,

$$\hat{v}_Q^y(x) := \frac{2\lambda}{1-\alpha}\,\Delta C(Q;P^{\mathrm{tar}})\,\frac{x}{\|x\|}\,\mathbb{1}_{\{\|x\|>y\}}, \tag{67}$$

depends on the target distribution only through $\mathrm{CVaR}_\alpha^{P^{\mathrm{tar}},g}$. We show that the finite-sample estimation error of $\mathrm{CVaR}_\alpha^{P^{\mathrm{tar}},g}$ does not depend on the dimension d, beginning with the following corollary.

Corollary D.12. *Assume that $\mathbb{E}_{P^{\mathrm{tar}}}[g^q] < \infty$ for some $q > 1$. Then, for a constant $C(q)$ independent of the dimension d of the target distribution P^{tar}, we have*

$$\mathbb{E}_{(P^{\mathrm{tar}})^{\otimes N}}\left|\mathrm{CVaR}_\alpha^{(\widehat{P^{\mathrm{tar}}})_N,g} - \mathrm{CVaR}_\alpha^{P^{\mathrm{tar}},g}\right| \le \frac{C(q)}{1-\alpha}(N^{-\frac{1}{2}} + N^{-\frac{(q-1)}{q}}). \tag{68}$$

If $q = 2$, the bound holds with q replaced by any $q' \in (1, 2)$.

Proof. Almost identical to the proof of Lemma B.5 and hence is omitted. ■

By Equation 68, the number of target samples required to estimate $\hat{v}_Q^y$ to a prescribed accuracy does not grow with d. By Equation 67, $\hat{v}_Q^y$ is active only on $\{\|x\| > y\}$, where its magnitude is $\frac{2\lambda}{1-\alpha}|\Delta C(Q; P^{\mathrm{tar}})|$; the choice of $y \in T(Q)$ therefore determines the region on which $\hat{v}_Q^y$ is active, not its magnitude. This magnitude is nonzero whenever $\Delta C(Q; P^{\mathrm{tar}}) \neq 0$, irrespective of the number of samples in the tail region, where the component $-\nabla_x \phi^*$ from D_{KL}^L may vanish due to sample scarcity. Since $\hat{v}_Q^y$ is proportional to $\Delta C(Q; P^{\mathrm{tar}})$, it is large when the tail is inadequately captured and vanishes as the CVaR discrepancy closes. The flow Equation 66 motivates the particle algorithm stated in Algorithm 4.

Activation region of the velocity contribution from CVaR-penalization The two endpoints $\mathrm{VaR}_\alpha^{\widehat{Q}_k,g}$ and $\overline{\mathrm{VaR}}_\alpha^{\widehat{Q}_k,g}$ of $T(\widehat{Q}_k)$ yield the largest and the smallest activation regions of $\hat{v}_Q^y$, respectively, and are computed exactly from the order statistics $g_{(1)} \le \cdots \le g_{(M)}$ of the risk values:

$$\mathrm{VaR}_\alpha^{\widehat{Q}_k,g} = g_{(\lceil \alpha M \rceil)}, \qquad \overline{\mathrm{VaR}}_\alpha^{\widehat{Q}_k,g} = g_{(\lfloor \alpha M \rfloor + 1)}. \tag{69}$$

If $\alpha M \notin \mathbb{Z}$, the two endpoints coincide. If $\alpha M \in \mathbb{Z}$, then $\mathrm{VaR}_\alpha^{\widehat{Q}_k,g} \le \overline{\mathrm{VaR}}_\alpha^{\widehat{Q}_k,g}$, with strict inequality precisely when $g_{(\alpha M)} < g_{(\alpha M+1)}$.

The update scheme Algorithm 4 follows the update scheme Equation 10 of CVaR-GPA, with the smoothed tail statistics replaced by their unsmoothed counterparts, computed via Algorithm 3, and with the velocity component induced by the CVaR-penalization replaced by

$$b_k^{(i)} = \frac{2\Delta_k}{1-\alpha}\frac{Y_k^{(i)}}{\|Y_k^{(i)}\|}\mathbb{1}_{\{\|Y_k^{(i)}\| > y_k\}}, \qquad y_k \in \left\{\mathrm{VaR}_\alpha^{\widehat{Q}_k,g}, \overline{\mathrm{VaR}}_\alpha^{\widehat{Q}_k,g}\right\},$$
$$\Delta_k = \mathrm{CVaR}_\alpha^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_\alpha^{\widehat{Q}_k,g}. \tag{70}$$

Algorithm 3 Empirical tail statistics

Require: samples $\{Z_i\}_{i=1}^m$, $g(\cdot) = \|\cdot\|$: the risk function, α: level

1: $g_i \leftarrow g(Z_i)$; sort ascending $g_{(1)} \le \cdots \le g_{(m)}$

2: $\mathrm{VaR} \leftarrow g_{(\lceil \alpha m \rceil)}, \quad \overline{\mathrm{VaR}} \leftarrow g_{(\lfloor \alpha m \rfloor + 1)}$

3: $\mathrm{CVaR} \leftarrow \mathrm{VaR} + \dfrac{1}{(1-\alpha)\,m}\displaystyle\sum_{i=\lceil \alpha m \rceil + 1}^{m}\big(g_{(i)} - \mathrm{VaR}\big)$ ▷ empirical form of Equation 50

4: **return** $(\mathrm{VaR},\ \overline{\mathrm{VaR}},\ \mathrm{CVaR})$

As for CVaR-GPA, the only target-dependent inputs of Algorithm 4 are the target samples $\{X^{(j)}\}$, so the algorithm is tail-agnostic.

Algorithm 4 CVaR-GPA using subgradients

Require: Hyperparameters α, λ, L, Δt; risk function $g(\cdot) = \|\cdot\|$; number of pre-trained particles M; number of target particles N; kinetic energy threshold ϵ; neural network approximation Γ_L^{NN}; number of network updates per iteration N_ϕ.

Result: Fine-tuned particles $\{Y_k^{(i)}\}_{i=1}^M$ at termination.

Sample $\{X^{(j)}\}_{j=1}^N \sim P^{\mathrm{tar}}$, a batch of samples from the target distribution.
Sample $\{Y_0^{(i)}\}_{i=1}^M \sim P^{\mathrm{pre}}$, a batch of samples from the pre-trained distribution.
Compute the target tail statistic $\mathrm{CVaR}_\alpha^{\widehat{P^{\mathrm{tar}}},g}$ via Algorithm 3 using $\{X^{(j)}\}_{j=1}^N$.
$k \leftarrow 0,\ \mathcal{K} \leftarrow 2\epsilon$
while $\mathcal{K} > \epsilon$ **do**
  $\phi_k^* \leftarrow \arg\max_{\phi \in \Gamma_L^{\mathrm{NN}}} \left\{ \frac{1}{M}\sum_i \phi(Y_k^{(i)}) - \log\left(\frac{1}{N}\sum_j e^{\phi(X^{(j)})}\right)\right\}$ ▷ N_ϕ steps
  $(\mathrm{VaR}_\alpha^{\widehat{Q}_k,g}, \overline{\mathrm{VaR}}_\alpha^{\widehat{Q}_k,g}, \mathrm{CVaR}_\alpha^{\widehat{Q}_k,g})$ via Algorithm 3 using $\{Y_k^{(i)}\}_{i=1}^M$.
  $y_k \leftarrow \mathrm{VaR}_\alpha^{\widehat{Q}_k,g}$ or $\overline{\mathrm{VaR}}_\alpha^{\widehat{Q}_k,g}$
  $\Delta_k \leftarrow \mathrm{CVaR}_\alpha^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_\alpha^{\widehat{Q}_k,g}$
  for $i = 1, \dots, M$ **do**
    $v_k^{(i)} \leftarrow -\nabla_x \phi_k^*(Y_k^{(i)}) + \dfrac{2\lambda}{1-\alpha}\,\Delta_k\,\dfrac{Y_k^{(i)}}{\|Y_k^{(i)}\|}\,\mathbb{1}\{\|Y_k^{(i)}\| > y_k\}$ ▷ Equation 70
  end for
  $Y_{k+1}^{(i)} \leftarrow Y_k^{(i)} + \Delta t\, v_k^{(i)};\quad \mathcal{K} \leftarrow \frac{1}{2M}\sum_i \|v_k^{(i)}\|^2$
  $k \leftarrow k+1$
end while
return $\{Y_k^{(i)}\}_{i=1}^M$

D.4 Using CVaR vs smoothed CVaR for penalization

Penalization via the unsmoothed CVaR is well posed: $\mathcal{F}$ is a divergence, its variational subgradients are characterized by Theorem D.7, and each subgradient Φ_Q^y, $y \in T(Q)$, defines a velocity field (Theorem D.8). The resulting flow, however, has three drawbacks, all of which stem from the indicator $\mathbb{1}_{\{g>y\}}$ in the subgradients $h^y = \frac{(g-y)^+}{1-\alpha}$.

1. *Subgradient selection.* Whenever $T(Q)$ is not a singleton and $\Delta C(Q; P^{\mathrm{tar}}) \neq 0$, the loss functional does not determine the velocity field: a level $y \in T(Q)$ must be selected, and different selections activate the CVaR velocity component on different regions. For empirical measures with $\alpha M \in \mathbb{Z}$, this is the generic situation (see Equation 69).
2. *Discontinuous velocity field.* The velocity field v_Q^y has a jump of magnitude $\frac{2\lambda}{1-\alpha}|\Delta C(Q; P^{\mathrm{tar}})|\,\|\nabla_x g\|$ across the threshold set $\{x : g(x) = y\}$ (Remark D.10).
3. *Unstable learning.* In Algorithm 4, the velocity component $\lambda b_k^{(i)}$, with $b_k^{(i)}$ as in Equation 70, assigns the full displacement $\frac{2\lambda}{1-\alpha}|\Delta_k|$ to the particles above y_k and none to those below it. Since both y_k and the particles move from one iteration to the next, a particle close to the threshold may cross it back and forth, so that the displacement it receives alternates between zero and its full magnitude.

Smoothing removes all three drawbacks while preserving the focus of the penalty on the tail. Replacing u^+ by the softplus $r_h(u)$ replaces the indicator $\mathbb{1}_{\{u>0\}}$, the derivative of u^+ for $u \neq 0$, by its smooth surrogate $r_h'(u) = \sigma(u/h)$, which converges to $\mathbb{1}_{\{u>0\}}$ as $h \downarrow 0$ for every $u \neq 0$. With $g = \|\cdot\|$, the two velocity components induced by the penalization are

$$b_k^{(i)} = \frac{2\Delta_k}{1-\alpha}\,\mathbb{1}_{\{\|Y_k^{(i)}\|>y_k\}}\,\frac{Y_k^{(i)}}{\|Y_k^{(i)}\|}, \qquad b_{k,h}^{(i)} = \frac{2\Delta_{k,h}}{1-\alpha}\,\sigma\Big(\frac{\|Y_k^{(i)}\| - q_{\alpha,h}^{\widehat{Q}_k,g}}{h}\Big)\,\frac{Y_k^{(i)}}{\|Y_k^{(i)}\|}, \tag{71}$$

where $\Delta_{k,h} = \mathrm{CVaR}_{\alpha,h}^{\widehat{P^{\mathrm{tar}}},g} - \mathrm{CVaR}_{\alpha,h}^{\widehat{Q}_k,g}$. Consequently:

1. $F^{Q,g}_{\alpha,h}$ is strictly convex with a unique minimizer $q^{Q,g}_{\alpha,h}$ (Proposition B.4), and $\mathcal{F}$ has a variational derivative at every $Q \in \mathcal{Q}^g$ (Theorem 2.2); no selection is required.
2. The jump of the CVaR velocity component across $\{x : g(x) = y\}$ is replaced by a smooth transition of width $O(h)$ around $\{x : g(x) = q^{Q,g}_{\alpha,h}\}$, and the weights $\sigma\big((\|Y^{(i)}_k\| - q^{\widehat{Q}_k,g}_{\alpha,h})/h\big)$ depend continuously on the particle positions.
3. The penalty remains focused on the tail. The weight $\sigma\big((g - q^{Q,g}_{\alpha,h})/h\big)$ is close to 1 above the level $q^{Q,g}_{\alpha,h}$ and close to 0 below it, and its total mass equals $1 - \alpha$ by Equation 14. Moreover, since r_h is convex and nondecreasing, $\mathrm{CVaR}^{Q,g}_{\alpha,h}$, viewed as a functional of the random variable $g(X)$, $X \sim Q$, is monotone, translation equivariant and convex, that is, a convex risk measure (Kouri & Surowiec, 2016). Finally, since $0 < r_h(u) - u^+ = h\log(1 + e^{-|u|/h}) \le h\log 2$, evaluating $F^{Q,g}_{\alpha,h}$ at any $y \in T(Q)$ and $F^{Q,g}_{\alpha}$ at $q^{Q,g}_{\alpha,h}$ gives

$$0 < \mathrm{CVaR}^{Q,g}_{\alpha,h} - \mathrm{CVaR}^{Q,g}_{\alpha} \le \frac{h\log 2}{1-\alpha}, \tag{72}$$

so the smoothed CVaR converges to the CVaR as $h \downarrow 0$, uniformly in $Q \in \mathcal{Q}^g$.

D.5 Proofs for the results concerning CVaR-penalized Wasserstein gradient flows with Clarke's generalized subgradients

This subsection collects the proofs of the results stated above, together with the auxiliary lemmas they require.

Lemma D.2 *Let $J(Q)$ be locally Lipschitz continuous with respect to the Wasserstein-1 metric on $\mathcal{P}(\mathbb{R}^d)$. Then for any admissible perturbation ρ, the Clarke generalized directional derivative $J^\circ(Q;\rho)$ exists and is finite.*

Proof. Let $J(Q)$ be locally Lipschitz continuous with respect to the Wasserstein-1 metric on $\mathcal{P}(\mathbb{R}^d)$. Let $\mathrm{Lip}(J, Q)$ denote the Lipschitz constant of the functional J in a neighborhood of Q. Then for an admissible perturbation ρ, we have

$$\begin{aligned}
|J^\circ(Q;\rho)| &= \left| \overline{\lim_{Q' \to Q,\ t \downarrow 0}} \frac{J(Q' + t\rho) - J(Q')}{t} \right| \\
&\le \overline{\lim_{Q' \to Q,\ t \downarrow 0}} \mathrm{Lip}(J, Q) \frac{\mathcal{W}_1(Q' + t\rho, Q')}{t} \\
&= \overline{\lim_{Q' \to Q,\ t \downarrow 0}} \mathrm{Lip}(J, Q) \frac{\sup_{\psi \in \Gamma_1} \int \psi \, d(Q' + t\rho - Q')}{t} \\
&= \mathrm{Lip}(J, Q) \sup_{\psi \in \Gamma_1} \int \psi \, d\rho \\
&= \mathrm{Lip}(J, Q) \sup_{\psi \in \Gamma_1} \int [\psi - \psi(0)] d\rho \quad (\text{since} \int d\rho = 0) \\
&\le \mathrm{Lip}(J, Q) \sup_{\psi \in \Gamma_1} \int [\psi - \psi(0)] d|\rho| \\
&\le \mathrm{Lip}(J, Q) \int \|x\| d|\rho| < \infty,
\end{aligned}$$

where the second equality follows from the Kantorovich–Rubinstein duality (Kantorovich & Rubinstein, 1958),

$$\sup_{\psi \in \Gamma_1} \int \psi \, d(Q_1 - Q_2) = \mathcal{W}_1(Q_1, Q_2).$$

■

Proposition D.3. *Assume that g is $\mathrm{Lip}(g)$-Lipschitz continuous and Assumption B.1 holds. Then the functional $Q \mapsto \mathcal{F}(Q; P^{\mathrm{tar}})$ is locally Lipschitz continuous on $\mathcal{Q}^g$ with respect to $\mathcal{W}_1$.*

Proof. Fix $y \in \mathbb{R}$. Since $t \mapsto t^+$ is 1-Lipschitz and g is $\mathrm{Lip}(g)$-Lipschitz, $x \mapsto (g(x) - y)^+$ is $\mathrm{Lip}(g)$-Lipschitz, uniformly in y. Thus we have

$$\begin{aligned}\left|F_\alpha^{Q_1,g}(y) - F_\alpha^{Q_2,g}(y)\right| &= \frac{1}{1-\alpha}\left|\mathbb{E}_{Q_1}[(g-y)^+] - \mathbb{E}_{Q_2}[(g-y)^+]\right| \\ &\le \frac{1}{1-\alpha} \sup_{\phi \in \Gamma_{\mathrm{Lip}(g)}} \int \phi \, d(Q_1 - Q_2) \\ &= \frac{\mathrm{Lip}(g)}{1-\alpha} \sup_{\psi \in \Gamma_1} \int \psi \, d(Q_1 - Q_2) \\ &= \frac{\mathrm{Lip}(g)}{1-\alpha} \mathcal{W}_1(Q_1, Q_2),\end{aligned}$$

for every y, uniformly. Here the last equality follows from the Kantorovich–Rubinstein duality (Kantorovich & Rubinstein, 1958)

$$\sup_{\psi \in \Gamma_1} \int \psi \, d(Q_1 - Q_2) = \mathcal{W}_1(Q_1, Q_2).$$

Thus it follows that

$$\begin{aligned}\left|\mathrm{CVaR}_\alpha^{Q_1,g} - \mathrm{CVaR}_\alpha^{Q_2,g}\right| &= \left|\inf_{y \in \mathbb{R}} F_\alpha^{Q_1,g}(y) - \inf_{y \in \mathbb{R}} F_\alpha^{Q_2,g}(y)\right| \\ &\le \sup_{y \in \mathbb{R}} \left|F_\alpha^{Q_1,g}(y) - F_\alpha^{Q_2,g}(y)\right| \\ &\le \frac{\mathrm{Lip}(g)}{1-\alpha} \mathcal{W}_1(Q_1, Q_2). \end{aligned} \tag{73}$$

Fix $Q_0 \in \mathcal{Q}^g$ and $R > 0$, and let $Q_1, Q_2 \in B_R(Q_0)$, so that $\mathcal{W}_1(Q_1, Q_0) \le R$ and $\mathcal{W}_1(Q_2, Q_0) \le R$. By Equation 73, for $i = 1, 2$

$$\left|\Delta C(Q_i; P^{\mathrm{tar}}) - \Delta C(Q_0; P^{\mathrm{tar}})\right| \;\le\; \frac{\mathrm{Lip}(g)}{1-\alpha} \mathcal{W}_1(Q_i, Q_0) \;\le\; \frac{\mathrm{Lip}(g)}{1-\alpha} R. \tag{74}$$

Hence

$$\begin{aligned}|\Delta C(Q_i; P^{\mathrm{tar}})| &\le |\Delta C(Q_0; P^{\mathrm{tar}})| + \left|\Delta C(Q_i; P^{\mathrm{tar}}) - \Delta C(Q_0; P^{\mathrm{tar}})\right| \\ &\le |\Delta C(Q_0; P^{\mathrm{tar}})| + \frac{\mathrm{Lip}(g)}{1-\alpha} R := M_R\end{aligned}$$

Then it readily follows that

$$\begin{aligned}&\left|\left(\Delta C(Q_1; P^{\mathrm{tar}})\right)^2 - \left(\Delta C(Q_2; P^{\mathrm{tar}})\right)^2\right| \\ =&\left|\Delta C(Q_1; P^{\mathrm{tar}}) + \Delta C(Q_2; P^{\mathrm{tar}})\right| \left|\Delta C(Q_1; P^{\mathrm{tar}}) - \Delta C(Q_2; P^{\mathrm{tar}})\right| \\ \le&\left(\left|\Delta C(Q_1; P^{\mathrm{tar}})\right| + \left|\Delta C(Q_2; P^{\mathrm{tar}})\right|\right) \left|\Delta C(Q_1; P^{\mathrm{tar}}) - \Delta C(Q_2; P^{\mathrm{tar}})\right| \\ \le&2M_R \left|\Delta C(Q_1; P^{\mathrm{tar}}) - \Delta C(Q_2; P^{\mathrm{tar}})\right| \\ \le&2M_R \frac{\mathrm{Lip}(g)}{1-\alpha} \mathcal{W}_1(Q_1, Q_2); \quad \text{by Equation 73.}\end{aligned}$$

Let $A^{Q_1}(\phi) := \mathbb{E}_{Q_1}[\phi] - \log \mathbb{E}_{P^{\mathrm{tar}}}[e^{\phi}]$ and $A^{Q_2}(\phi) := \mathbb{E}_{Q_2}[\phi] - \log \mathbb{E}_{P^{\mathrm{tar}}}[e^{\phi}]$ for $\phi \in \Gamma_L$. Then, by the variational representation Equation 3, we have

$$\begin{aligned} D_{\mathrm{KL}}^L(Q_1\|P^{\mathrm{tar}}) - D_{\mathrm{KL}}^L(Q_2\|P^{\mathrm{tar}}) &= \sup_{\phi\in\Gamma_L} A^{Q_1}(\phi) - \sup_{\phi\in\Gamma_L} A^{Q_2}(\phi) \\ &\le \sup_{\phi\in\Gamma_L} \left\{A^{Q_1}(\phi) - A^{Q_2}(\phi)\right\} \\ &= \sup_{\phi\in\Gamma_L} \left\{\mathbb{E}_{Q_1}[\phi] - \mathbb{E}_{Q_2}[\phi]\right\} \\ &= \sup_{\phi\in\Gamma_L} \int \phi\, d(Q_1 - Q_2) \\ &= L \sup_{\psi\in\Gamma_1} \int \psi\, d(Q_1 - Q_2) \\ &= L\mathcal{W}_1(Q_1, Q_2), \end{aligned}$$

where the last equality follows from the Kantorovich–Rubinstein duality. A similar argument yields $D_{\mathrm{KL}}^L(Q_2\|P^{\mathrm{tar}}) - D_{\mathrm{KL}}^L(Q_1\|P^{\mathrm{tar}}) \le L\mathcal{W}_1(Q_1, Q_2)$. Hence we obtain

$$\left|D_{\mathrm{KL}}^L(Q_1\|P^{\mathrm{tar}}) - D_{\mathrm{KL}}^L(Q_2\|P^{\mathrm{tar}})\right| \le L\,\mathcal{W}_1(Q_1, Q_2). \tag{75}$$

Then we have

$$\begin{aligned} &\left|\mathcal{F}(Q_1; P^{\mathrm{tar}}) - \mathcal{F}(Q_2; P^{\mathrm{tar}})\right| \\ \le &\left|D_{\mathrm{KL}}^L(Q_1\|P^{\mathrm{tar}}) - D_{\mathrm{KL}}^L(Q_2\|P^{\mathrm{tar}})\right| + \lambda\left|(\Delta C(Q_1; P^{\mathrm{tar}}))^2 - (\Delta C(Q_2; P^{\mathrm{tar}}))^2\right| \\ \le &L\,\mathcal{W}_1(Q_1, Q_2) + 2\lambda M_R \frac{\mathrm{Lip}(g)}{1-\alpha}\,\mathcal{W}_1(Q_1, Q_2) \\ = &K_R\,\mathcal{W}_1(Q_1, Q_2) \end{aligned}$$

where $K_R = L + 2\lambda M_R \frac{\mathrm{Lip}(g)}{1-\alpha}$. Since $Q_0 \in \mathcal{Q}^g$ and $R > 0$ were arbitrary, every $Q \in \mathcal{Q}^g$ has a $\mathcal{W}_1$-neighborhood on which $\mathcal{F}(\cdot\,; P^{\mathrm{tar}})$ is Lipschitz, i.e., $\mathcal{F}(\cdot\,; P^{\mathrm{tar}})$ is locally Lipschitz continuous with respect to $\mathcal{W}_1$. ■

Lemma D.4 *Let $J(Q)$ be locally Lipschitz continuous with respect to the Wasserstein-1 metric on $\mathcal{P}(\mathbb{R}^d)$. Define the right/left one-sided directional derivatives of J as:*

$$D_\rho^{\pm} J := \lim_{\epsilon\to 0^{\pm}} \frac{1}{\epsilon}\big(J(Q^\epsilon) - J(Q)\big).$$

Assume that for every admissible perturbation ρ at Q, the one-sided directional derivatives $D_\rho^{\pm} J$ exist and $\int \xi\, d\rho$ lies in the closed interval with endpoints $D_\rho^- J$ and $D_\rho^+ J$. Then ξ is a variational subgradient of J at Q.

Proof. Let ρ be an admissible perturbation at Q. The right/left one-sided derivatives of J at Q are defined by

$$D_\rho^{\pm} J := \lim_{\epsilon\to 0^{\pm}} \frac{1}{\epsilon}\big(J(Q^\epsilon) - J(Q)\big). \tag{76}$$

Restricting the limit superior in the definition of $J^\circ(Q;\rho)$ to the constant family $Q' = Q$ gives

$$J^\circ(Q;\rho) \;\ge\; \lim_{t\downarrow 0} \frac{J(Q + t\rho) - J(Q)}{t} \;=\; D_\rho^+ J.$$

Similarly, we may instead take the family $Q'_t := Q - t\rho$, which lies in $\mathcal{P}(\mathbb{R}^d)$ for all small $t > 0$ by the admissibility of $-\rho$ and converges to Q in $\mathcal{W}_1$ as $t \downarrow 0$. Along this family

$$\frac{J(Q'_t + t\rho) - J(Q'_t)}{t} \;=\; \frac{J(Q) - J(Q - t\rho)}{t} \;=\; \left.\frac{J(Q + s\rho) - J(Q)}{s}\right|_{s=-t} \;\xrightarrow[t\downarrow 0]{}\; D_\rho^- J,$$

whence $J^\circ(Q;\rho) \ge D_\rho^- J$ as well. Therefore $J^\circ(Q;\rho) \ge \max\{D_\rho^+ J,\, D_\rho^- J\} \ge \int \xi\, d\rho$, and $\xi \in \partial J(Q)$, i.e. ξ is a variational subgradient of $J(Q)$. ■

We have the following auxiliary lemma.

Lemma D.13. *Assume that g is Lipschitz continuous and $Q \in \mathcal{Q}^g$. Let ρ be right-admissible at Q. Then there exists $\epsilon_0 > 0$ such that*

1. *The function $(\epsilon, y) \mapsto F_\alpha^{Q^\epsilon, g}(y)$ is jointly continuous in $[0, \epsilon_0] \times \mathbb{R}$.*
2. *The set $T := \bigcup_{\epsilon \in [0,\epsilon_0]} T(Q^\epsilon)$ is compact.*

Proof. (1) Denote by $\ell_\rho(y) := \frac{1}{1-\alpha} \int (g(x) - y)^+ \, d\rho(x)$ for any $y \in \mathbb{R}$ and $\|\rho\| = \int d|\rho|$. For all $y \in \mathbb{R}$, $|\ell_\rho(y)| \le \frac{1}{1-\alpha} \left[\int g \, d|\rho| + |y| \, \|\rho\| \right] < \infty$ since ρ is right-admissible and $\int g \, d|\rho| < \infty$ by Equation 56. Let ϵ_0 be a given threshold and let $(\epsilon, y) \in [0, \epsilon_0] \times \mathbb{R}$ be arbitrary. Let $\tilde{\epsilon} > 0$ and $\delta < \frac{\tilde{\epsilon}}{E_0}$ where $E_0 := \frac{2-\alpha}{1-\alpha} + |\ell_\rho(y)| + \frac{\epsilon_0 \|\rho\|}{1-\alpha}$. For $\epsilon' \in [0, \epsilon_0]$ let $\|(y, \epsilon) - (y', \epsilon')\| < \delta$. Since the function $t \mapsto t^+$ is 1-Lipschitz continuous, $|(g(x) - y)^+ - (g(x) - y')^+| \le |y - y'|$ for all $x \in \mathbb{R}^d$. Hence, for any signed measure π, we have the following

$$\left| \int \left[(g(x) - y)^+ - (g(x) - y')^+ \right] d\pi \right| \le \int |y - y'| d|\pi| = |y - y'| \int d|\pi| \tag{77}$$

$$\begin{aligned}
&|F_\alpha^{Q^\epsilon, g}(y) - F_\alpha^{Q^{\epsilon'}, g}(y')| \\
=&\left| y - y' + \frac{1}{1-\alpha} \int \left[(g(x) - y)^+ - (g(x) - y')^+ \right] dQ + \epsilon \ell_\rho(y) - \epsilon' \ell_\rho(y') \right| \\
\le& |y - y'| + \frac{1}{1-\alpha} |y - y'| + |\epsilon \ell_\rho(y) - \epsilon' \ell_\rho(y')| \, ; \text{ (from Equation 77 with } \pi = Q) \\
=& \frac{2-\alpha}{1-\alpha} |y - y'| + |\epsilon \ell_\rho(y) - \epsilon' \ell_\rho(y')| \\
\le& \frac{2-\alpha}{1-\alpha} |y - y'| + |\epsilon \ell_\rho(y) - \epsilon' \ell_\rho(y)| + |\epsilon' \ell_\rho(y) - \epsilon' \ell_\rho(y')| \\
\le& \frac{2-\alpha}{1-\alpha} |y - y'| + |\ell_\rho(y)| \, |\epsilon - \epsilon'| + \epsilon' \, |\ell_\rho(y) - \ell_\rho(y')| \\
\le& \frac{2-\alpha}{1-\alpha} |y - y'| + |\ell_\rho(y)| \, |\epsilon - \epsilon'| + \epsilon_0 \, |\ell_\rho(y) - \ell_\rho(y')| \\
\le& \frac{2-\alpha}{1-\alpha} |y - y'| + |\ell_\rho(y)| \, |\epsilon - \epsilon'| + \frac{\epsilon_0 \, \|\rho\|}{1-\alpha} \, |y - y'| \, ; \text{ (from Equation 77 with } \pi = \rho) \\
\le& \left(\frac{2-\alpha}{1-\alpha} + |\ell_\rho(y)| + \frac{\epsilon_0 \, \|\rho\|}{1-\alpha} \right) \delta = E_0 \delta < \tilde{\epsilon}.
\end{aligned}$$

Hence, $F_\alpha^{Q^\epsilon, g}(y)$ is jointly continuous in $[0, \epsilon_0] \times \mathbb{R}$.

(2) For any $c \in \mathbb{R}$, $|\Psi^{Q^\epsilon, g}(c) - \Psi^{Q, g}(c)| \le \epsilon \|\rho\|$. In the trivial case $\|\rho\| = 0$, $T = T(Q)$, and the result readily follows. Assume $\|\rho\| > 0$. Fix $\zeta \in (0, \frac{1}{2} \min\{\alpha, 1 - \alpha\})$ and $a < b$ with $\Psi^{Q,g}(a) \le \alpha - 2\zeta$, $\Psi^{Q,g}(b) \ge \alpha + 2\zeta$; set $\epsilon_0 = \min\{\zeta / \|\rho\|, \bar{\epsilon}\}$, where $\bar{\epsilon} > 0$ is an admissibility bound ($Q^\epsilon \in \mathcal{P}(\mathbb{R}^d)$ for $\epsilon \in [0, \bar{\epsilon}]$). Then for all $\epsilon \in [0, \epsilon_0]$ we have

$$\Psi^{Q^\epsilon, g}(a) \le \alpha - \zeta < \alpha, \qquad \Psi^{Q^\epsilon, g}(b) \ge \alpha + \zeta > \alpha. \tag{78}$$

Thus, by definition of $\mathrm{VaR}_\alpha^{Q^\epsilon, g}$ and $\overline{\mathrm{VaR}}_\alpha^{Q^\epsilon, g}$ we have that $a \le \mathrm{VaR}_\alpha^{Q^\epsilon, g} \le \overline{\mathrm{VaR}}_\alpha^{Q^\epsilon, g} \le b$. Therefore, $T(Q^\epsilon) = [\mathrm{VaR}_\alpha^{Q^\epsilon, g}, \overline{\mathrm{VaR}}_\alpha^{Q^\epsilon, g}] \subseteq [a, b]$ for every $\epsilon \in [0, \epsilon_0]$. Thus $T := \bigcup_{\epsilon \in [0,\epsilon_0]} T(Q^\epsilon)$ is a bounded set.

Let $y_n \in T$ be a sequence that converges to $\hat{y}$. For each n, there exists $\epsilon_n \in [0, \epsilon_0]$ such that $y_n \in T(Q^{\epsilon_n})$. By the compactness of the interval $[0, \epsilon_0]$, there exists a subsequence ϵ_{n_k} that converges to $\epsilon^* \in [0, \epsilon_0]$. Since $(\epsilon, y) \mapsto F_\alpha^{Q^\epsilon, g}(y)$ is jointly continuous in $[0, \epsilon_0] \times \mathbb{R}$ for any threshold ϵ_0, we have the following for any $z \in \mathbb{R}$

$$F_\alpha^{Q^{\epsilon^*}, g}(\hat{y}) = \lim_{k \to \infty} F_\alpha^{Q^{\epsilon_{n_k}}, g}(y_{n_k}) \le \lim_{k \to \infty} F_\alpha^{Q^{\epsilon_{n_k}}, g}(z) = F_\alpha^{Q^{\epsilon^*}, g}(z). \tag{79}$$

That is, $\hat{y} \in T(Q^{\epsilon^*}) \subseteq T$. Thus, T is closed, which proves the result. ∎

Theorem D.5 *Assume that g is Lipschitz continuous and Q has finite first moment (hence $Q \in \mathcal{Q}^g$). For $y \in \mathbb{R}$, define $h^y(x) := \frac{(g(x)-y)^+}{1-\alpha}$. Then we have the following:*

1. (Right derivative) *For every right-admissible perturbation ρ, the right directional derivative exists and is given by*
$$D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g} = \lim_{\epsilon \to 0^+} \frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} = \min_{y \in T(Q)} \int h^y \, d\rho. \tag{57}$$
2. (Left derivative) *For every left-admissible perturbation ρ, the left directional derivative exists and is given by*
$$D_\rho^- \mathrm{CVaR}_\alpha^{Q,g} = \lim_{\epsilon \to 0^-} \frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} = \max_{y \in T(Q)} \int h^y \, d\rho. \tag{58}$$
3. (Variational subgradients) *Let $y \in T(Q)$. Then, for any admissible ρ, $\int h^y \, d\rho$ lies in $[D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g}, D_\rho^- \mathrm{CVaR}_\alpha^{Q,g}]$. Consequently, for each $y \in T(Q)$, h^y is a variational subgradient of $\mathrm{CVaR}_\alpha^{Q,g}$.*
4. (Two-sided derivative) *The two-sided (Gateaux) derivative for an admissible ρ exists if and only if $y \mapsto \int h^y \, d\rho$ is constant on $T(Q)$; this holds for every admissible ρ if and only if $T(Q)$ is a singleton set, i.e. $\mathrm{VaR}_\alpha^{Q,g} = \overline{\mathrm{VaR}}_\alpha^{Q,g}$, in which case $\mathrm{CVaR}_\alpha^{Q,g}$ is Gateaux differentiable at Q with the first variation*
$$\frac{\delta \mathrm{CVaR}_\alpha^{Q,g}}{\delta Q} = h^{\mathrm{VaR}_\alpha^{Q,g}} = \frac{(g - \mathrm{VaR}_\alpha^{Q,g})^+}{1-\alpha}.$$

Proof. For any perturbation ρ, we define $\ell_\rho(y) := \frac{1}{1-\alpha} \int (g-y)^+ d\rho = \int h^y \, d\rho$.

(1) Let ρ be right-admissible and let $\epsilon_0 > 0$ be given by Lemma D.13. From $Q^\epsilon = Q + \epsilon\rho$, we have
$$F_\alpha^{Q^\epsilon,g}(y) = F_\alpha^{Q,g}(y) + \epsilon \, \ell_\rho(y). \tag{80}$$
Thus, it follows that
$$\begin{aligned} \mathrm{CVaR}_\alpha^{Q^\epsilon,g} &\le \inf_{y \in T(Q)} F_\alpha^{Q^\epsilon,g}(y) \\ &= \inf_{y \in T(Q)} \{F_\alpha^{Q,g}(y) + \epsilon \ell_\rho(y)\} \\ &= \mathrm{CVaR}_\alpha^{Q,g} + \epsilon \min_{y \in T(Q)} \ell_\rho(y). \end{aligned} \tag{81}$$
Hence, we have
$$\overline{\lim_{\epsilon \to 0^+}} \frac{(\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g})}{\epsilon} \le \min_{y \in T(Q)} \ell_\rho(y) = \min_{y \in T(Q)} \int h^y \, d\rho. \tag{82}$$
Pick $y_\epsilon \in T(Q^\epsilon)$. Using Equation 80 and $F_\alpha^{Q,g}(y_\epsilon) \ge \mathrm{CVaR}_\alpha^{Q,g}$, we have
$$\frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} = \frac{F_\alpha^{Q,g}(y_\epsilon) - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} + \ell_\rho(y_\epsilon) \ge \ell_\rho(y_\epsilon). \tag{83}$$
Fix $\epsilon_n \downarrow 0$; by compactness of T pass to a subsequence with $y_{\epsilon_n} \to \bar{y} \in T$. Since $y \mapsto \ell_\rho(y)$ is continuous on $\mathbb{R}$, there exists $M > 0$ such that $\sup_{y \in T} |\ell_\rho(y)| \le M$. Thus, for any $y^* \in \mathbb{R}$, we have
$$\begin{aligned} F_\alpha^{Q,g}(y_{\epsilon_n}) &= F_\alpha^{Q^{\epsilon_n},g}(y_{\epsilon_n}) - \epsilon_n \ell_\rho(y_{\epsilon_n}) \\ &\le F_\alpha^{Q^{\epsilon_n},g}(y^*) - \epsilon_n \ell_\rho(y_{\epsilon_n}) \\ &= F_\alpha^{Q,g}(y^*) + \epsilon_n \ell_\rho(y^*) - \epsilon_n \ell_\rho(y_{\epsilon_n}) \\ &\le F_\alpha^{Q,g}(y^*) + M\epsilon_n + |\ell_\rho(y^*)| \, \epsilon_n. \end{aligned}$$

Letting $\epsilon_n \to 0$, and using the continuity of $F_\alpha^{Q,g}(\cdot)$, we obtain that $\lim_{n\to\infty} F_\alpha^{Q,g}(y_{\epsilon_n}) = F_\alpha^{Q,g}(\bar{y}) \le F_\alpha^{Q,g}(y^*)$ for all $y^* \in \mathbb{R}$. Thus $\bar{y} \in T(Q)$.

Thus, by Equation 83, and the continuity of $\ell_\rho(\cdot)$, we have

$$\underline{\lim}_{n\to\infty} \frac{\mathrm{CVaR}_\alpha^{Q^{\epsilon_n},g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon_n} \ge \lim_{n\to\infty} \ell_\rho(y_{\epsilon_n}) = \ell_\rho(\bar{y}) \ge \min_{y\in T(Q)} \ell_\rho(y) = \min_{y\in T(Q)} \int h^y \, d\rho. \tag{84}$$

Thus, we can conclude using Equation 82 and Equation 84 that for any sequence $\epsilon_n \to 0^+$ in the interval $(0, \epsilon_0]$, there exists a subsequence ϵ_{n_k} such that

$$\lim_{\epsilon_{n_k}\to 0^+} \frac{\mathrm{CVaR}_\alpha^{Q^{\epsilon_{n_k}},g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon_{n_k}} = \min_{y\in T(Q)} \int h^y \, d\rho. \tag{85}$$

Hence,

$$\lim_{\epsilon\to 0^+} \frac{\mathrm{CVaR}_\alpha^{Q^{\epsilon},g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} = \min_{y\in T(Q)} \int h^y \, d\rho, \tag{86}$$

proving Equation 57.

(2) Let ρ be left-admissible. Let $\epsilon = -\delta$, $\delta > 0$. Thus $Q^\epsilon = Q + \epsilon\rho = Q - \delta\rho$. We have

$$\begin{aligned} D_\rho^- \mathrm{CVaR}_\alpha^{Q,g} &= \lim_{\epsilon\to 0^-} \frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\epsilon} \\ &= -\lim_{\delta\to 0^+} \frac{\mathrm{CVaR}_\alpha^{Q-\delta\rho,g} - \mathrm{CVaR}_\alpha^{Q,g}}{\delta} \\ &= -\min_{y\in T(Q)} \int h^y \, d(-\rho) \\ &= \max_{y\in T(Q)} \int h^y \, d\rho. \end{aligned} \tag{87}$$

(3) Since $y \mapsto \ell_\rho(y)$ is continuous and $T(Q)$ is a compact interval, $\int h^y d\rho = \ell_\rho(y)$ ranges over the interval between $D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g}$ and $D_\rho^- \mathrm{CVaR}_\alpha^{Q,g}$, attaining both extremes. Thus, by Lemma D.4, it follows that the family of functions $\{h^y; y \in T(Q)\}$ are variational subgradients of $\mathrm{CVaR}_\alpha^{Q,g}$.

(4) Let ρ be a perturbation such that ρ is admissible at Q. By parts (1) and (2), both one-sided directional derivatives exist and are given by

$$D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g} = \min_{y\in T(Q)} \ell_\rho(y), \qquad D_\rho^- \mathrm{CVaR}_\alpha^{Q,g} = \max_{y\in T(Q)} \ell_\rho(y).$$

The two-sided (Gateaux) derivative in direction ρ exists if and only if these one-sided derivatives coincide, that is, if and only if

$$\min_{y\in T(Q)} \ell_\rho(y) = \max_{y\in T(Q)} \ell_\rho(y).$$

That is, the two-sided (Gateaux) derivative in direction ρ exists if and only if $y \mapsto \ell_\rho(y)$ is constant on $T(Q)$.

Assume that $T(Q)$ is a singleton set, i.e. $\mathrm{VaR}_\alpha^{Q,g} = \overline{\mathrm{VaR}}_\alpha^{Q,g}$, so that $T(Q) = \{\mathrm{VaR}_\alpha^{Q,g}\}$. Then $y \mapsto \ell_\rho(y)$ is constant on $T(Q)$ for every admissible ρ, and

$$D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g} = D_\rho^- \mathrm{CVaR}_\alpha^{Q,g} = \ell_\rho\big(\mathrm{VaR}_\alpha^{Q,g}\big) = \int h^{\mathrm{VaR}_\alpha^{Q,g}} \, d\rho.$$

The two-sided derivative therefore exists for every admissible perturbation; that is, $\mathrm{CVaR}_\alpha^{Q,g}$ is Gateaux differentiable at Q with first variation

$$\frac{\delta \mathrm{CVaR}_\alpha^{Q,g}}{\delta Q} = h^{\mathrm{VaR}_\alpha^{Q,g}} = \frac{(g - \mathrm{VaR}_\alpha^{Q,g})^+}{1-\alpha}.$$

Conversely, assume that $\mathrm{VaR}_\alpha^{Q,g} < \overline{\mathrm{VaR}}_\alpha^{Q,g}$. Define $k(x) = \big(g(x) - \mathrm{VaR}_\alpha^{Q,g}\big)^+ - \big(g(x) - \overline{\mathrm{VaR}}_\alpha^{Q,g}\big)^+$. Define $\varphi_k := k - \mathbb{E}_Q[k]$. Since $0 \le k \le \overline{\mathrm{VaR}}_\alpha^{Q,g} - \mathrm{VaR}_\alpha^{Q,g}$, both k and φ_k are bounded. Define the perturbation measure $\rho := \varphi_k\, Q$. We verify that ρ is admissible at Q. First, $\int d\rho = \int \varphi_k\, dQ = \mathbb{E}_Q[k] - \mathbb{E}_Q[k] = 0$ and $\int d|\rho| = \int |\varphi_k|\, dQ \le \|\varphi_k\|_\infty < \infty$. Second, for $|\epsilon| \le \|\varphi_k\|_\infty^{-1}$ we have $1 + \epsilon\varphi_k \ge 0$ point-wise, so $Q^\epsilon = Q + \epsilon\rho = (1 + \epsilon\varphi_k)\, Q$ is a non-negative measure of total mass 1, i.e. $Q^\epsilon \in \mathcal{P}(\mathbb{R}^d)$ for all $|\epsilon| \le \|\varphi_k\|_\infty^{-1}$. Third, we have

$$\int \|x\|\, d|\rho| = \int \|x\|\, |\varphi_k|\, dQ \le \|\varphi_k\|_\infty \int \|x\|\, dQ(x), \tag{88}$$

and since φ_k is bounded and Q has a finite first moment, we have $\int \|x\|\, d|\rho| < \infty$. Hence, ρ is admissible. Since $\rho = \varphi_k Q$,

$$\begin{aligned}
\ell_\rho\big(\mathrm{VaR}_\alpha^{Q,g}\big) - \ell_\rho\big(\overline{\mathrm{VaR}}_\alpha^{Q,g}\big) &= \frac{1}{1-\alpha} \int \Big[\big(g - \mathrm{VaR}_\alpha^{Q,g}\big)^+ - \big(g - \overline{\mathrm{VaR}}_\alpha^{Q,g}\big)^+\Big]\, d\rho \\
&= \frac{1}{1-\alpha} \int k\, \varphi_k\, dQ \\
&= \frac{1}{1-\alpha}\, \mathbb{E}_Q\big[k\,(k - \mathbb{E}_Q[k])\big] \\
&= \frac{\mathrm{Var}_Q(k)}{1-\alpha},
\end{aligned}$$

where $\mathrm{Var}_Q(k)$ denotes the variance of k under the probability measure Q. Since $\mathrm{VaR}_\alpha^{Q,g} < \overline{\mathrm{VaR}}_\alpha^{Q,g}$, the CDF $\Psi^{Q,g}$ satisfies $\Psi^{Q,g}(c) = \alpha$ for all $c \in [\mathrm{VaR}_\alpha^{Q,g}, \overline{\mathrm{VaR}}_\alpha^{Q,g})$. Consequently, $Q(\{k = 0\}) = Q(g \le \mathrm{VaR}_\alpha^{Q,g}) = \Psi^{Q,g}(\mathrm{VaR}_\alpha^{Q,g}) = \alpha$ and $Q(\{k > 0\}) = 1 - \alpha > 0$. Since the probability of $k > 0$ and the probability of $k = 0$ are strictly positive under the probability measure Q, $\mathrm{Var}_Q(k) > 0$. Thus we have

$$\ell_\rho\big(\mathrm{VaR}_\alpha^{Q,g}\big) - \ell_\rho\big(\overline{\mathrm{VaR}}_\alpha^{Q,g}\big) = \frac{\mathrm{Var}_Q(k)}{1-\alpha} > 0. \tag{89}$$

Thus ℓ_ρ is non-constant on $T(Q)$, and hence

$$D_\rho^+ \mathrm{CVaR}_\alpha^{Q,g} = \min_{y \in T(Q)} \ell_\rho(y) < \max_{y \in T(Q)} \ell_\rho(y) = D_\rho^- \mathrm{CVaR}_\alpha^{Q,g}.$$

The two-sided derivative fails to exist for the perturbation ρ. We conclude that the two-sided derivative exists for every admissible perturbation if and only if $T(Q)$ is a singleton set. ■

Theorem D.7 *Assume that g is Lipschitz continuous and Assumption B.1 holds. Let ϕ^* denote the maximizer of Equation 3. For each $y \in T(Q)$, define the potential function*

$$\Phi_Q^y := \phi^* - 2\lambda\, \Delta C(Q; P^{\mathrm{tar}})\, h^y = \phi^* - \frac{2\lambda}{1-\alpha}\, \Delta C(Q; P^{\mathrm{tar}})\, (g - y)^+. \tag{59}$$

Then the following hold.

1. *For every right-admissible perturbation ρ, the right directional derivative exists and is given by*

$$D_\rho^+ \mathcal{F}(Q; P^{\mathrm{tar}}) = \int \phi^*\, d\rho - 2\lambda\, \Delta C(Q; P^{\mathrm{tar}}) \min_{y \in T(Q)} \int h^y d\rho. \tag{60}$$

   For every left-admissible perturbation ρ, the left directional derivative exists and is given by

$$D_\rho^- \mathcal{F}(Q; P^{\mathrm{tar}}) = \int \phi^*\, d\rho - 2\lambda\, \Delta C(Q; P^{\mathrm{tar}}) \max_{y \in T(Q)} \int h^y d\rho. \tag{61}$$

2. *For any admissible ρ, the range of $\int \Phi_Q^y d\rho$ for $y \in T(Q)$ is the compact interval between $D_\rho^- \mathcal{F}(Q; P^{\mathrm{tar}})$ and $D_\rho^+ \mathcal{F}(Q; P^{\mathrm{tar}})$. Consequently, the family of potential functions $\{\Phi_Q^y; y \in T(Q)\}$ are subgradients of $\mathcal{F}(Q; P^{\mathrm{tar}})$.*

3. *Let ρ be admissible. The two-sided (Gateaux) derivative of $\mathcal{F}$ for the perturbation ρ exists if and only if $\Delta C(Q;P^{\mathrm{tar}})=0$ or $y\mapsto\int h^y\,d\rho$ is constant on $T(Q)$. The two-sided (Gateaux) derivative of $\mathcal{F}$ exists for every admissible perturbation if and only if $\Delta C(Q;P^{\mathrm{tar}})=0$ or $T(Q)$ is a singleton set; in that case, $\mathcal{F}(\cdot;P^{\mathrm{tar}})$ is Gateaux differentiable at Q with first variational derivative*

$$\frac{\delta\mathcal{F}(Q;P^{\mathrm{tar}})}{\delta Q}=\Phi_Q^{\mathrm{VaR}_\alpha^{Q,g}}.$$

Proof. (1) For a right-admissible perturbation ρ, we have that

$$\begin{aligned}
D_\rho^+(\Delta C(Q;P^{\mathrm{tar}}))^2 &= \lim_{\epsilon\to 0^+}\frac{(\Delta C(Q^\epsilon;P^{\mathrm{tar}}))^2-(\Delta C(Q;P^{\mathrm{tar}}))^2}{\epsilon}\\
&= \lim_{\epsilon\to 0^+}\left[\mathrm{CVaR}_\alpha^{Q^\epsilon,g}+\mathrm{CVaR}_\alpha^{Q,g}-2\mathrm{CVaR}_\alpha^{P^{\mathrm{tar}},g}\right]\frac{\mathrm{CVaR}_\alpha^{Q^\epsilon,g}-\mathrm{CVaR}_\alpha^{Q,g}}{\epsilon}\\
&= -2\Delta C(Q;P^{\mathrm{tar}})\,D_\rho^+\mathrm{CVaR}_\alpha^{Q,g}\\
&= -2\Delta C(Q;P^{\mathrm{tar}})\min_{y\in T(Q)}\int h^y\,d\rho. \qquad (90)
\end{aligned}$$

We recall that the variational derivative of the divergence $D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})$ is given by

$$\frac{\delta D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})}{\delta Q}(x)=\phi^*(x), \qquad (91)$$

where ϕ^* is the maximizer of Equation 3. That is, for any right-admissible ρ, $D_\rho^+D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})=\int\phi^*d\rho$. Since $D_\rho^+\mathcal{F}(Q;P^{\mathrm{tar}})=D_\rho^+D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})+\lambda D_\rho^+(\Delta C(Q;P^{\mathrm{tar}}))^2$, we readily obtain Equation 60.

Now let ρ be a left-admissible perturbation. Then we have $D_\rho^-D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})=\int\phi^*d\rho$ and $D_\rho^-(\Delta C(Q;P^{\mathrm{tar}}))^2=-2\Delta C(Q;P^{\mathrm{tar}})\max_{y\in T(Q)}\int h^y\,d\rho$ (similar to the derivation of Equation 90). Since $D_\rho^-\mathcal{F}(Q;P^{\mathrm{tar}})=D_\rho^-D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})+\lambda D_\rho^-(\Delta C(Q;P^{\mathrm{tar}}))^2$, we readily obtain Equation 61.

(2) Since $y\mapsto\int\Phi_Q^y\,d\rho$ is continuous and $T(Q)$ is a compact interval, $\int\Phi_Q^y\,d\rho$ ranges over the interval between $D_\rho^+\mathcal{F}(Q;P^{\mathrm{tar}})$ and $D_\rho^-\mathcal{F}(Q;P^{\mathrm{tar}})$, attaining both extremes. Thus, by Lemma D.4, it follows that the family of potential functions $\{\Phi_Q^y;y\in T(Q)\}$ are subgradients of $\mathcal{F}(Q;P^{\mathrm{tar}})$.

(3) For any admissible perturbation ρ, we have

$$D_\rho^-\mathcal{F}(Q;P^{\mathrm{tar}})-D_\rho^+\mathcal{F}(Q;P^{\mathrm{tar}})=2\lambda\,\Delta C(Q;P^{\mathrm{tar}})\Big(\min_{y\in T(Q)}\int h^y\,d\rho-\max_{y\in T(Q)}\int h^y\,d\rho\Big).$$

The two-sided derivative for perturbation ρ exists if and only if $\Delta C(Q;P^{\mathrm{tar}})=0$ or $\int h^yd\rho$ is constant on $T(Q)$. As proven in Theorem D.5, $\int h^yd\rho$ is constant on $T(Q)$ for every admissible ρ, if and only if $T(Q)=\{\mathrm{VaR}_\alpha^{Q,g}\}$ is a singleton set. If $T(Q)$ is the singleton set $\{\mathrm{VaR}_\alpha^{Q,g}\}$, we have

$$D_\rho^-\mathcal{F}(Q;P^{\mathrm{tar}})=D_\rho^+\mathcal{F}(Q;P^{\mathrm{tar}})=\int\phi^*d\rho-2\lambda\,\Delta C(Q;P^{\mathrm{tar}})\int h^{\mathrm{VaR}_\alpha^{Q,g}}d\rho,$$

for every admissible ρ. That is, $\mathcal{F}(\cdot;P^{\mathrm{tar}})$ is Gateaux differentiable at Q with first variational derivative

$$\frac{\delta\mathcal{F}(Q;P^{\mathrm{tar}})}{\delta Q}=\phi^*-2\lambda\,\Delta C(Q;P^{\mathrm{tar}})\,h^{\mathrm{VaR}_\alpha^{Q,g}}=\Phi_Q^{\mathrm{VaR}_\alpha^{Q,g}}.$$

■

Theorem D.8 *Assume that g is Lipschitz continuous, $g \in C^1$ for $g(x) \neq 0$, and Assumption B.1 holds. For each $y \in T(Q)$, choosing the potential function Φ_Q^y induces the velocity field $v_Q^y = -\nabla_x \Phi_Q^y$, and we have*

$$v_Q^y(x) = \begin{cases} \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})} + \underbrace{\dfrac{2\lambda}{1-\alpha}\,\Delta C(Q;P^{\mathrm{tar}})\,\nabla_x g(x)}_{\text{from CVaR-penalization}}, & g(x) > y, \\ \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})}, & g(x) \le y. \end{cases} \tag{62}$$

If $T(Q)$ is a singleton set, the velocity field degenerates to $v_Q^{\mathrm{VaR}_\alpha^{Q,g}}$.

Proof of Theorem D.8. The result readily follows from taking the gradient of $\Phi_Q^y(x)$ with respect to x for $g(x) > y \geq 0$. Since $(g(x)-y)^+$ is not differentiable at $g(x) = y$, the velocity field is not defined at $g(x) = y$. One can adopt the convention that $v^y(x) \in \{-\nabla_x \phi^*(x) + s\left(\frac{2\lambda}{(1-\alpha)}\Delta C(Q;P^{\mathrm{tar}})\nabla_x g(x)\right); s \in [0,1]\}$ for $g(x) = y$. In our work, we define $v^y(x) = -\nabla_x \phi^*(x)$ for $g(x) = y$, yielding Equation 62. ■

Corollary D.9 *Assume that g is Lipschitz continuous, $g \in C^1$ for $g(x) \neq 0$, and Assumption B.1 holds. For every $y \in T(Q)$, the velocity field v_Q^y in Equation 62 satisfies, for almost every $x \in \mathbb{R}^d$,*

$$\left\|v_Q^y(x)\right\| \;\le\; L + \frac{2\lambda}{1-\alpha}\,\mathrm{Lip}(g)\,\left|\Delta C(Q;P^{\mathrm{tar}})\right|. \tag{63}$$

Moreover, the CVaR discrepancy is controlled by the loss $\mathcal{F}(Q;P^{\mathrm{tar}})$, so we have

$$\left\|v_Q^y(x)\right\| \;\le\; L + \frac{2\,\mathrm{Lip}(g)}{1-\alpha}\,\sqrt{\lambda\,\mathcal{F}(Q;P^{\mathrm{tar}})}, \tag{64}$$

for almost every $x \in \mathbb{R}^d$. In particular, the velocity field is bounded, and the CVaR-penalization component vanishes as $\mathcal{F}(Q;P^{\mathrm{tar}}) \to 0$, i.e. as $Q \to P^{\mathrm{tar}}$.

Proof of Corollary D.9. By Equation 62, for almost every x, we have

$$\left\|v_Q^y(x)\right\| \le \left\|\nabla_x \phi^*(x)\right\| + \frac{2\lambda}{1-\alpha}\left|\Delta C(Q;P^{\mathrm{tar}})\right|\left\|\nabla_x g(x)\right\|.$$

Since $\phi^* \in \Gamma_L$ is L-Lipschitz, $\|\nabla_x \phi^*\| \leq L$ almost everywhere; since g is $\mathrm{Lip}(g)$-Lipschitz, $\|\nabla_x g\| \leq \mathrm{Lip}(g)$. This proves Equation 63.

Since $D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}}) \geq 0$ we have

$$\lambda\left(\Delta C(Q;P^{\mathrm{tar}})\right)^2 \le D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}}) + \lambda\left(\Delta C(Q;P^{\mathrm{tar}})\right)^2 = \mathcal{F}(Q;P^{\mathrm{tar}})\,.$$

That is $|\Delta C(Q;P^{\mathrm{tar}})| \leq \sqrt{\mathcal{F}(Q;P^{\mathrm{tar}})/\lambda}$; this combined with Equation 63, proves Equation 64. ■

Corollary D.11 *Let Assumption B.1 hold and let $g(x) = \|x\|$ and $\mathbb{E}_Q[g] < \infty$. For each $y \in T(Q)$, choosing the potential function Φ_Q^y induces the velocity field $v_Q^y = -\nabla_x \Phi_Q^y$, and we have*

$$v_Q^y(x) = \begin{cases} \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})} + \underbrace{\dfrac{2\lambda}{1-\alpha}\,\Delta C(Q;P^{\mathrm{tar}})\,\dfrac{x}{\|x\|}}_{\text{from CVaR-penalization}}, & \|x\| > y, \\ \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\mathrm{KL}}^L(Q\|P^{\mathrm{tar}})}, & \|x\| \le y. \end{cases} \tag{65}$$

Proof of Corollary D.11. Immediate from Theorem D.8 with $g(x) = \|x\|$ and $\nabla_x\|x\| = x/\|x\|$ for $x \neq 0$. ■