EXACT AND INEXACT MARGINAL PENALTIES: WHEN DOES THE WASSERSTEIN AUTOENCODER RECOVER THE WASSERSTEIN DISTANCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Primal optimal transport (OT) minimizes a transport cost over couplings with prescribed marginals. Parameterizing a coupling by a conditional distribution enforces one marginal by construction and leaves the other as an explicit constraint. The Wasserstein autoencoder relaxes this remaining constraint into a divergence penalty. We characterize when the resulting penalized objective recovers the OT cost at a finite coefficient λ. On bounded domains and under an active constraint, f-divergences with f differentiable at 1, as well as squared maximum mean discrepancy, strictly underestimate the cost for every finite coefficient. For f satisfying a generalized Pinsker inequality, the optimal-value gap is $\Theta(1/\lambda)$ as $\lambda \to \infty$. In contrast, the square roots of such f-divergences and the 1-Wasserstein distance recover the OT value and global minimizers above explicit finite thresholds, and for square roots of the Kullback-Leibler and chi-square divergences exactness extends to unbounded priors. The dichotomy follows from comparing reconstruction descent with penalty growth in the Wasserstein space of joint distributions. Uniform exactness over a decoder class further yields statistical guarantees for learning the pushforward model. Simulations with computable ground truth illustrate the dichotomy, while CelebA experiments compare latent matching and local image redundancy across penalty classes.

1 INTRODUCTION

The Wasserstein distance remains informative when two distributions are concentrated near low-dimensional manifolds, rendering it a natural loss for generative modeling (Arjovsky et al., 2017; Genevay et al., 2018). However, computing it between continuous distributions is far harder than between finitely-supported ones, where Sinkhorn iterations apply (Cuturi, 2013). For fixed continuous measures, prior approaches either parameterize Kantorovich dual potentials (Genevay et al., 2016; Buzun et al., 2024) or learn transport plans and flows directly from samples (Korotin et al., 2023; Xu et al., 2025); optimizing a generator outside a dual solver gives the min–max formulation of Wasserstein GANs (WGANs; Arjovsky et al., 2017). On benchmarks with analytically known solutions, however, Korotin et al. (2021) find that many evaluated quadratic-cost solvers do not faithfully recover optimal maps, while Korotin et al. (2022) find that the evaluated WGAN dual solvers need not accurately compute W_1. These findings concern approximation error in particular neural optimal transport (OT) solvers. We study a distinct, logically prior source of error: when a marginal constraint in primal OT is relaxed into a divergence penalty, does the resulting penalized objective still recover the original OT cost?

The primal identity underlying the Wasserstein autoencoder (WAE; Tolstikhin et al., 2018) provides a setting in which this formulation-level question can be posed exactly. A decoder $g : \mathcal{Z} \to \mathcal{X}$ pushes a prior P_Z on $\mathcal{Z} \subset \mathbb{R}^d$ forward to a model distribution $P_G := g_\sharp P_Z$ on $\mathcal{X} \subset \mathbb{R}^m$. Let $\mathcal{Q}^*$ be the set of regular conditional distributions $Q_{Z|X}$ of Z given $X \sim P_X$, called encoders, and let $Q_{Z|X}P_X$ be the induced joint distribution of (Z, X). For a cost $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, Tolstikhin et al. (2018, Theorem 1) show that

$$W_c(P_X, g_\sharp P_Z) = \inf_{Q_{Z|X} \in \mathcal{C}} \mathbb{E}_{(Z,X)\sim Q_{Z|X}P_X} c(X, g(Z)), \quad \mathcal{C} = \{Q_{Z|X} \in \mathcal{Q}^* : Q_Z = P_Z\}, \qquad \text{(O)}$$

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.

Table 1: Finite-coefficient dichotomy for marginal penalties. For $c = d_{\mathcal{X}}^p$ and a γ-Lipschitz decoder, the cost contributes the same $L_p(g) = p\,\mathrm{diam}(\mathcal{X})^{p-1}\max\{1,\gamma\}$ to every row (Lemma 3.2). Inexact rows assume an active constraint (Remark 3.6). All results are population-level, and Section 4 applies to every exact row.

Marginal Penalty	Penalty / prior assumptions	Conclusion
D_f	f differentiable at 1	Inexact for all finite λ
	generalized Pinsker	Gap $\Theta(1/\lambda)$ as $\lambda \to \infty$
MMD_k^2	Bounded continuous characteristic k	Inexact; gap $\Omega(1/\lambda)$
$\sqrt{D_f}$	Bounded $\mathcal{Z}$; generalized Pinsker	Exact if $\lambda > L_p(g)\,\mathrm{diam}(\mathcal{Z})/\sqrt{2f''(1)}$
W_1	Finite first moments	Exact if $\lambda > L_p(g)$
$\sqrt{\mathrm{KL}}$	$W_1^2(Q, P_Z) \le 2\tau\mathrm{KL}(Q, P_Z)$	Exact if $\lambda > L_p(g)\sqrt{2\tau}$
$\sqrt{\chi^2}$	$0 < \sigma^2 = \mathbb{E}_{P_Z}\|Z\|^2 < \infty$	Exact if $\lambda > L_p(g)\sigma$

where $Q_Z := \pi_{1\sharp}(Q_{Z|X}P_X)$ is the aggregate posterior and $\pi_1 : (z, x) \mapsto z$. For a fixed decoder, (O) is primal OT between P_Z and P_X with lifted cost $(z, x) \mapsto c(x, g(z))$. The conditional-law parameterization anchors the first marginal at P_X, leaving $Q_Z = P_Z$ as the only explicit constraint. When $\mathcal{Z} = \mathcal{X}$ and $g = \mathrm{id}$, (O) reduces to ordinary OT between P_Z and P_X. Further minimizing (O) in g over some set $\mathcal{G}$ of measurable functions yields a joint minimization formulation for learning a generative model at the population level, where the constraint $Q_Z = P_Z$ ensures that generation from the prior is consistent with the latent codes on which the decoder was trained. In practice, this remaining marginal constraint is replaced by a penalty,

$$V_\lambda(g) := \inf_{Q_{Z|X} \in \mathcal{Q}^*} \{\mathbb{E}_{(Z,X)\sim Q_{Z|X}P_X} c(X, g(Z)) + \lambda D(Q_Z, P_Z)\}, \tag{P}$$

with a divergence $D(\cdot, \cdot)$ and a coefficient $\lambda > 0$, which permits stochastic gradient training (Bottou et al., 2018). Common choices for D include the squared maximum mean discrepancy (MMD, Gretton et al., 2012) and f-divergences such as Jensen-Shannon (JS). Thus, (P) is a one-sided marginal penalization of primal OT; only the latent marginal is penalized. We ask whether a finite coefficient recovers both the OT cost and the constrained problem's global minimizers, i.e., $V_\lambda(g) = W_c(P_X, g_\sharp P_Z)$ for a finite λ for a common $Q_{Z|X} \in \mathcal{C}$. Failure of either constitutes a formulation bias that precedes encoder or critic restriction and stochastic optimization.

We answer this question with exact penalty theory. In metric spaces, exactness follows from two error bounds relative to the feasible set (Dolgopolik, 2016), which leave open which marginal penalties satisfy them and what occurs when they fail. We settle both for the marginal-penalized primal formulation induced by the WAE identity, by recasting (O) as an optimization over the 1-Wasserstein space of joint distributions on $\mathcal{Z} \times \mathcal{X}$. Let r denote the W_1 distance to the feasible set. Reconstruction can fall below the OT cost at most linearly in r, so any marginal penalty that globally dominates r at a linear rate is exact above a finite coefficient. Conversely, under an active constraint, the constructed perturbation path decreases reconstruction linearly in r. Along this path, f-divergences with f differentiable at 1 and squared MMD with a bounded continuous kernel grow as $o(r)$, so the penalized value remains strictly below the OT cost at every finite coefficient. When $f \in C^2$ with $f''(1) > 0$, and also for characteristic squared MMD, the gap admits a lower bound inversely proportional to the coefficient. Moreover, f-divergences satisfying the generalized Pinsker condition (Gilardoni, 2010) gives the exact $\Theta(1/\lambda)$ rate. This class includes the standard f-divergences such as Kullback-Leibler (KL), JS, and χ^2. In contrast, square roots of these f-divergences and the 1-Wasserstein distance dominate r linearly under the stated assumptions and are exact above explicit finite thresholds (Table 1 and Figure 1). The results for $\sqrt{\mathrm{KL}}$ and $\sqrt{\chi^2}$ also cover unbounded latent priors.

The fixed-decoder results characterize the inner marginal-penalized OT problem. With a uniform decoder Lipschitz bound, exactness holds simultaneously over a decoder class, so minimizing the outer marginal-penalized objective is equivalent to minimizing the OT cost. This yields generalization bounds with no complexity term for the decoder class and distribution-estimation guarantees governed by the intrinsic data dimension. Fixed-decoder simulations compare reconstruction estimates with computable OT ground truth and illustrate both sides of the finite-coefficient dichotomy. CelebA experiments then study joint encoder–decoder training. The square-root penalties improve latent coverage under the logit-normal prior with no degradation in FID. All theoretical statements concern

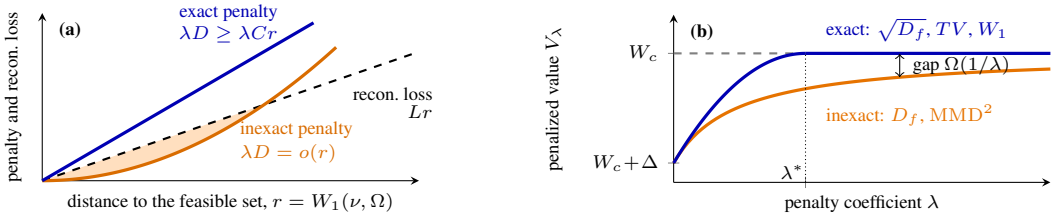


Figure 1: The mechanism behind Table 1. (a) Let $r = W_1(\nu, \Omega)$ denote distance to feasibility. Reconstruction can fall below W_c by at most Lr. Under an active constraint, the path constructed in Proposition 3.4 realizes a reconstruction decrease of order r. If $D = o(r)$ along this path, no finite λ offsets the decrease; the shaded region marks $V_\lambda < W_c$. In contrast, $D \geq Cr$ offsets every such decrease when $\lambda > L/C$. (b) Exact penalties yield $V_\lambda = W_c$ above a finite threshold, whereas $W_c - V_\lambda = \Omega(1/\lambda)$ at every finite λ for $D = D_f$ with $f \in C^2$ and $f''(1) > 0$, or for $D = \mathrm{MMD}^2$ with a characteristic kernel (Corollary 3.8 and Proposition 3.4).

population objectives and global minimizers; finite networks, variational divergence estimation, and optimization introduce additional errors beyond the formulation bias isolated here.

2 Background and Related Work

Transport and divergences. For $P, Q \in \mathcal{P}(\mathcal{X})$ and a cost c, the OT cost is $W_c(P, Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y)$, where $\Pi(P, Q)$ is the set of joint probability measures having P and Q as marginals. For $c = d_{\mathcal{X}}^p$ with a metric $d_{\mathcal{X}}$ and $p \geq 1$, $W_p := W_c^{1/p}$ is the p-Wasserstein distance on $\mathcal{P}_p(\mathcal{X}) = \{P : \mathbb{E}_P d_{\mathcal{X}}(x, x_0)^p < \infty\}$, and $\mathbb{W}_p(\mathcal{X}) = (\mathcal{P}_p(\mathcal{X}), W_p)$ is the p-Wasserstein space. Each W_p is weakly lower semicontinuous (l.s.c.), and $W_1(P, Q) = \sup_{\phi \in \mathcal{H}_1} \{\mathbb{E}_P \phi - \mathbb{E}_Q \phi\}$ over the set $\mathcal{H}_1$ of 1-Lipschitz functions.

Let $f : [0, \infty) \to (-\infty, \infty]$ be convex, finite on $(0, \infty)$, and continuous at 0, with $f(1) = 0$. Writing $P = P_{\mathrm{ac}} + P_\perp$ for the Lebesgue decomposition of P with respect to Q, the f-divergence is $D_f(P, Q) = \int f(dP_{\mathrm{ac}}/dQ)\, dQ + f'(\infty) P_\perp(\mathcal{X})$, where $f'(\infty) := \lim_{x \to \infty} f(x)/x$. The KL, JS, and χ^2 divergences correspond to $f(x) = x \log x$, $x \log x - (x + 1) \log \frac{x+1}{2}$, and $(x - 1)^2$. Every D_f is weakly l.s.c. For a bounded continuous kernel k with mean embedding $m_P := \int k(\cdot, z) dP(z)$, the squared MMD is $\mathrm{MMD}_k^2(P, Q) := \|m_P - m_Q\|^2$, and k is characteristic if $P \mapsto m_P$ is injective.

Exact penalty methods. Consider minimizing $F(x)$ over a feasible set $\Omega \subset A$. A penalty method minimizes $F(x) + \lambda \mathcal{D}(x)$ for a penalty $\mathcal{D} \geq 0$ that vanishes exactly on Ω. The penalty is *exact* if the penalized problem solves the constrained one at a *finite* λ, so that a single unconstrained problem suffices. Dolgopolik (2016) gives sufficient conditions in metric spaces.

Theorem 2.1 (Dolgopolik (2016))**.** *Let $(\mathcal{M}, d_{\mathcal{M}})$ be a metric space and $A \subset \mathcal{M}$. Suppose F attains its global minimum F^* over $\Omega \subset A$. Assume there exist constants $L, C > 0$ and $\alpha \in (0, 1]$ such that (i) $F(x) \geq F^* - L\{d_{\mathcal{M}}(x, \Omega)\}^\alpha$ for all $x \in A \setminus \Omega$, and (ii) $\mathcal{D}(x) \geq C\{d_{\mathcal{M}}(x, \Omega)\}^\alpha$ for all $x \in A$, where $d_{\mathcal{M}}(x, \Omega) := \inf_{y \in \Omega} d_{\mathcal{M}}(x, y)$. Then, for any $\lambda > L/C$, the penalty $\mathcal{D}(\cdot)$ is exact for the problem* $\min_{x \in \Omega} F(x)$.

Related work. Prior work has established two sufficient conditions under which (P) recovers (O). For the metric costs, Husain et al. (2019) show that the two problems have equal values for $D = D_f$ provided that $\lambda \geq \sup_{P'} \{W_1(P', P_G)/D_f(P', P_G)\}$. This supremum, however, is infinite when f is differentiable at 1. Indeed, for any fixed distribution $P' \neq P_G$ that is absolutely continuous with respect to P_G, the mixture path $P_\alpha = (1 - \alpha)P_G + \alpha P'$ yields $W_1(P_\alpha, P_G) = \alpha W_1(P', P_G)$, whereas $D_f(P_\alpha, P_G) = o(\alpha)$ as $\alpha \downarrow 0$ (Polyanskiy & Wu, 2025, Theorem 7.19). Consequently, this sufficient condition becomes vacuous for standard f-divergences and leaves open the existence of a finite exact coefficient; Section 3.2 resolves this question in the negative. The second condition is due to Patrini et al. (2020, Theorem 2.1). For $c = \|\cdot\|_p^p$ and a non-atomic P_X, they show that if λ

exceeds the Lipschitz constant of g, then

$$W_p(P_X, P_G) = \inf_{Q_{Z|X} \in \mathcal{Q}^*} \Big\{ \big[\mathbb{E}_{Q_{Z|X} P_X} \|X - g(Z)\|_p^p\big]^{1/p} + \lambda W_p(Q_Z, P_Z) \Big\}. \tag{1}$$

The p-th root biases stochastic gradients unless $p = 1$, so what is minimized in practice is an upper bound with a Sinkhorn proxy for $W_p(Q_Z, P_Z)$ rather than (O) or (P). In this work, Appendix C extends (1) using the framework of Section 3 without requiring non-atomicity, and Corollary 3.8 covers the p-th power cost that WAEs typically minimize. Statistical analyses of WAEs (Chakrabarty & Das, 2021; Chakrabarty et al., 2025; Chakraborty & Bartlett, 2024) are discussed in Section 4.2, and Appendix F provides a detailed comparison between each line of prior work and the contributions of this paper.

3 Exact and Inexact Marginal Penalties for Primal OT

For this section, the decoder g is fixed. We study whether penalizing the P_Z-marginal recovers the value and global minimizers of the primal OT problem (O). All statements concern the population objective and its global minimizers over the full class $\mathcal{Q}^*$ of encoders. Section 4 turns to learning g, and proofs are in Appendix B. A summary of the assumptions and results is provided in Appendix E.

3.1 Marginal-Penalized Primal OT on a Wasserstein Space

Throughout this paper, we assume that P_X and P_Z have finite first moments and endow $\mathcal{Z} \times \mathcal{X}$ with the sum of the marginal metrics, for which the first moment of a joint distribution is the sum of the first moments of its marginals. Every feasible distribution of (O), having marginals P_Z and P_X, therefore lies in the metric space $\mathbb{W}_1(\mathcal{Z} \times \mathcal{X})$. Fixing the X-marginal defines the domain $A = \{\mu \in \mathbb{W}_1(\mathcal{Z} \times \mathcal{X}) : \pi_{2\sharp}\mu = P_X\}$, with $\pi_2 : (z, x) \mapsto x$, and the feasible set of (O) is $\Omega := \{\mu \in A : \pi_{1\sharp}\mu = P_Z\}$. Problems (O) and (P) can then be equivalently recast as

$$\min_{\mu \in A} \mathcal{R}(\mu) \quad \text{subject to} \quad \mathcal{D}(\mu) = 0, \tag{O$'$}$$

$$\min_{\mu \in A} \{\mathcal{R}(\mu) + \lambda \mathcal{D}(\mu)\}, \tag{P$'$}$$

where $\mathcal{R}(\mu) = \mathbb{E}_\mu\, c(X, g(Z))$ and $\mathcal{D}(\mu) = D(\pi_{1\sharp}\mu, P_Z)$. Theorem 2.1 with $\alpha = 1$ yields the following sufficient conditions on $\mathcal{R}(\cdot)$ and $\mathcal{D}(\cdot)$ for exactness, where lower semicontinuity of the cost guarantees a minimizer of (O$'$).

Corollary 3.1. *Suppose that the reconstruction cost $(z, x) \mapsto c(x, g(z))$ is l.s.c., and let $\mathcal{R}^*$ denote the optimal value of problem (O$'$). Suppose that there exist positive constants L and C satisfying*

$$(i)\ \mathcal{R}(\nu) \geq \mathcal{R}^* - L \cdot W_1(\nu, \Omega), \quad \forall \nu \in A \setminus \Omega, \tag{2}$$

$$(ii)\ \mathcal{D}(\nu) \geq C \cdot W_1(\nu, \Omega), \quad \forall \nu \in A, \tag{3}$$

where $W_1(\nu, \Omega) := \inf_{\omega \in \Omega} W_1(\nu, \omega)$. Then $\mathcal{D}(\cdot)$ is exact for any $\lambda > L/C$.

Condition (2) depends only on the objective and is shared across all penalties, whereas condition (3) involves only the penalty. For the standard cost, condition (2) holds with an explicit constant.

Lemma 3.2. *Suppose that $\mathcal{X}$ is bounded, g is γ-Lipschitz, and $c = d_{\mathcal{X}}^p$ with $p \geq 1$. Then condition (2) holds with $L_p(g) = p\,\mathrm{diam}(\mathcal{X})^{p-1} \max\{1, \gamma\}$. For $p = 1$, the boundedness of $\mathcal{X}$ is not required.*

A second key identity, used on both sides of the dichotomy below, equates the distance to the feasible set with the latent mismatch:

$$W_1(\nu, \Omega) = W_1(\pi_{1\sharp}\nu, P_Z) \qquad \text{for all } \nu \in A. \tag{4}$$

This property holds because $\pi_{1\sharp}$ is 1-Lipschitz in the sum metric and the latent marginal of ν can be transported onto P_Z while holding its X-coordinate fixed; see Appendix B.3 for the proof. Condition (3) therefore imposes a requirement solely on the penalty, that it dominate $W_1(Q_Z, P_Z)$. Lemma 3.2 bounds the cost decrease by a linear function of this distance, and Section 3.2 establishes that, when the constraint is active, this decrease achieves that linear rate along a specific direction. Consequently, exactness hinges on whether the penalty grows at least linearly (Figure 1a).

3.2 FINITE-COEFFICIENT INEXACTNESS

Every weakly l.s.c. penalty recovers (O) in the limit as its coefficient diverges. This asymptotic fact does not determine whether a finite coefficient recovers the OT cost.

Theorem 3.3 (Convergence Theorem). *Let $\mathcal{Z} \times \mathcal{X}$ be compact, let $(z, x) \mapsto c(x, g(z))$ be continuous, and let $D(\cdot, \cdot)$ be weakly l.s.c. and vanish only on the diagonal. Let $\lambda_k \uparrow \infty$ and let μ_k minimize (P′) at $\lambda = \lambda_k$. Then $\{\mu_k\}$ has a weakly convergent subsequence, and every weak limit point of $\{\mu_k\}$ lies in Ω and solves (O′).*

This convergence guarantee encompasses D_f, W_p, and MMD^2 with a continuous characteristic kernel. To investigate the behavior at finite penalty scales, let $W_c := W_c(P_X, g_\sharp P_Z)$ and $V_\lambda \equiv V_\lambda(g)$. Then, $V_\lambda \leq W_c$ for every $\lambda > 0$, since any feasible encoder is admissible in (P). We show that D_f and MMD_k^2 strictly underestimate the OT cost. Under additional regularity, the D_f gap is $\Theta(1/\lambda)$, while for characteristic MMD_k^2 we prove an $\Omega(1/\lambda)$ lower bound.

Proposition 3.4 (Finite-coefficient inexactness). *Let $\mathcal{X}$ and $\mathcal{Z}$ be bounded with $\mathrm{supp}(P_Z) = \mathcal{Z}$, let $(x, z) \mapsto c(x, g(z))$ be continuous and bounded, and suppose that the constraint is* active*:*

$$\Delta := \mathbb{E}_{P_X}\Big[\inf_{z \in \mathcal{Z}} c(X, g(z))\Big] - W_c < 0. \tag{5}$$

If either $D = D_f$ with f differentiable at 1, or $D = \mathrm{MMD}_k^2$ with a bounded continuous kernel k, then $V_\lambda < W_c$ for every $\lambda > 0$.

Theorem 3.5 (Order-$1/\lambda$ inexactness gap). *Under the assumptions of Proposition 3.4, let $D = D_f$ with $f \in C^2(0, \infty)$ and $f''(1) > 0$. There exists $c_1 > 0$, independent of λ, such that $W_c - V_\lambda \geq \min\{c_1/\lambda, |\Delta|/8\}$ for all $\lambda > 0$. Furthermore, if $c = d_{\mathcal{X}}^p$, g is γ-Lipschitz, and f satisfies the generalized Pinsker condition of Appendix B.9, then $W_c - V_\lambda \leq (L_p(g)\ \mathrm{diam}(\mathcal{Z}))^2/(8f''(1)\lambda)$ for all $\lambda > 0$, with $L_p(g)$ from Lemma 3.2. If $D = \mathrm{MMD}_k^2$ with characteristic k, there exists $c_M > 0$ such that $W_c - V_\lambda \geq \min\{c_M/\lambda, |\Delta|/4\}$ for all $\lambda > 0$.*

Furthermore, any global penalized minimizer μ^λ of (P′) satisfies the non-asymptotic reconstruction bound:

$$\min\{c_1/\lambda, |\Delta|/8\} \leq W_c - \mathcal{R}(\mu^\lambda) \leq \big(L_p(g)\,\mathrm{diam}(\mathcal{Z})\big)^2/(2f''(1)\lambda) \tag{6}$$

This consequence on the reconstruction error is proved with Theorem 3.5.

The proofs proceed by perturbing a feasible, nearly optimal joint distribution $\bar{\mu} \in \Omega$ along the mixture segment $\mu_\alpha = (1 - \alpha)\bar{\mu} + \alpha\rho$ in A, where ρ is the joint distribution induced by the quantile encoder that restricts the prior to the θ-fraction ($\theta \in (0, 1)$) of latent codes minimizing the reconstruction cost $c(x, g(z))$, yielding a latent marginal $\pi_{1\sharp}\rho$ with density at most $1/\theta$ with respect to P_Z (Appendix B.5). Because the constraint is active, the cost decreases linearly in α. Differentiability of f at 1 makes $D_f = o(\alpha)$ along this path, which is enough for strict inexactness. Under the C^2 assumptions of Theorem 3.5, $D_f = \mathcal{O}(\alpha^2)$; MMD^2 is exactly quadratic in α. Balancing this trade-off yields the lower gap bounds in Theorem 3.5. The $o(\alpha)$ relation shows that condition (3) fails for every $C > 0$.

Remark 3.6 (Constraint activity). By Lemma B.5, $\Delta = V_0 - W_c$, so constraint activity signifies that dropping the penalty strictly decreases the optimal value. The equality $\Delta = 0$ holds if and only if some coupling $\pi \in \Pi(P_Z, P_X)$ is concentrated on $\{(z, x) : c(x, g(z)) = \inf_{z'} c(x, g(z'))\}$, that is, if assigning each x to an optimal code already reproduce the prior. In this degenerate case, $\lambda = 0$ is trivially exact, so $\Delta < 0$ is necessary for any inexactness statement. It holds, for instance, when the optimal code $T(x) \in \arg\min_z c(x, g(z))$ is unique P_X-almost everywhere and $T_\sharp P_X \neq P_Z$.

Does the gap matter? The gap vanishes as $\lambda \to \infty$, suggesting that one might simply increase λ until the bias is negligible. However, the constraint remains violated at every finite λ, and scaling up λ creates a regime where the penalty gradient dominates and optimization deteriorates. An exact penalty avoids both issues at an appropriate finite coefficient at the population optimum.

3.3 EXACTNESS ABOVE AN EXPLICIT THRESHOLD

Theorem 3.7. *Assume that $\mathcal{Z}$ is bounded. Then condition (3) holds (a) for $D = \sqrt{D_f}$ with $C = \sqrt{2f''(1)}/\mathrm{diam}(\mathcal{Z})$, if f satisfies the generalized Pinsker condition (Gilardoni, 2010) stated in Appendix B.9, (b) for $D = \mathrm{TV}$ with $C = 1/\mathrm{diam}(\mathcal{Z})$, and (c) for $D = W_1$ with $C = 1$.*

The divergences employed in the f-GAN (Nowozin et al., 2016), namely KL, reverse KL, JS, squared Hellinger, χ^2, and reverse χ^2, all satisfy the generalized Pinsker condition (Appendix B.9). The proof combines $\sqrt{D_f} \geq \sqrt{2f''(1)}\,\mathrm{TV}$, the transport inequality $W_1 \leq \mathrm{diam}(\mathcal{Z})\,\mathrm{TV}$, and the identity (4). For these divergences D_f itself is inexact by Proposition 3.4; hence, taking the square root is precisely what transforms these inexact penalties into exact ones. Combining Corollary 3.1, Lemma 3.2, and Theorem 3.7 yields our primary exactness guarantee.

Corollary 3.8 (Exactness)**.** *Suppose that $\mathcal{X}$ and $\mathcal{Z}$ are bounded, g is γ-Lipschitz, $c = d_{\mathcal{X}}^p$ with $p \geq 1$, and D is one of the penalties of Theorem 3.7, with constant C. Then for every $\lambda > \lambda^*(g) := L_p(g)/C$, the global minimizers of (P′) coincide with those of (O′). In particular, (P′) attains its infimum, and its value equals $W_p^p(P_X, g_\sharp P_Z)$.*

For $D = W_1$ and $p = 1$ the threshold is $\max\{1, \gamma\}$, which reduces to the Lipschitz threshold of Patrini et al. (2020) when $\gamma \geq 1$.

3.4 Exactness on Unbounded Latent Spaces

The boundedness of $\mathcal{Z}$ can be relaxed to a condition on the prior that covers standard Gaussian priors in practice. By Lemma 3.2, the boundedness of $\mathcal{X}$ is not required when $p = 1$.

$\sqrt{\mathrm{KL}}$ penalty. Suppose that P_Z satisfies the $T_1(\tau)$ transportation inequality (Villani, 2009, Theorem 22.10), namely $\sqrt{2\tau D_{KL}(Q, P_Z)} \geq W_1(Q, P_Z)$ for all Q. Combined with (4), this inequality establishes condition (3) for $\mathcal{D}(\nu) = \sqrt{D_{KL}(\pi_{1\sharp}\nu, P_Z)}$ with $C = 1/\sqrt{2\tau}$. This condition holds for Gaussian and, more generally, strongly log-concave priors on $\mathbb{R}^d$ (Villani, 2009).

$\sqrt{\chi^2}$ penalty. A second-moment condition suffices, which is strictly weaker than tail concentration.

Proposition 3.9. *Suppose that P_Z has a finite second moment and set $\sigma^2 := \mathbb{E}_{P_Z}\|Z\|^2 < \infty$. Then $\sigma\sqrt{\chi^2(Q, P_Z)} \geq W_1(Q, P_Z)$ for all Q. Consequently, the penalty $\mathcal{D}(\nu) = \sqrt{\chi^2(\pi_{1\sharp}\nu, P_Z)}$ satisfies condition (3) with $C = 1/\sigma$.*

4 Learning the Decoder: Statistical Consequences

Section 3.3 fixed the decoder and studied the inner penalized problem. We now optimize over a decoder class. Exactness turns this outer problem into minimizing the Wasserstein distance, so the statistical analysis can run on W_p directly. Throughout this section $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Z} \subset \mathbb{R}^d$.

4.1 Generalization Error

We quantify the intrinsic dimension of P_X as follows. Let $N_\varepsilon(S)$ be the minimal number of open balls of radius ε that cover $S \subset \mathcal{X}$, and let $N_\varepsilon(P, \eta) = \inf\{N_\varepsilon(S) : P(S) \geq 1 - \eta\}$ for $P \in \mathcal{P}(\mathcal{X})$.

Definition 4.1 (Wasserstein dimension (Weed & Bach, 2019))**.** Define the (ϵ, η)-dimension of a measure P as $d_\varepsilon(P, \eta) := \log N_\varepsilon(P, \eta)/(-\log \varepsilon)$. The upper p-Wasserstein dimension of P is $d_p^*(P) := \inf\{s \in (2p, \infty) : \limsup_{\varepsilon \to 0} d_\varepsilon(P, \varepsilon^{\frac{sp}{s-2p}}) \leq s\}$.

For the empirical measure $\hat{P}_X^n$ of n i.i.d. samples from P_X, it holds that $\mathbb{E}[W_p^p(P_X, \hat{P}_X^n)] = \mathcal{O}(n^{-p/s})$ for any $s > d_p^*(P_X)$ (Weed & Bach, 2019, Proposition 5). Let $V_{c,D,\lambda}(P_X, g_\sharp P_Z)$ denote the optimal value of (P), which is also the objective of the decoder, and write $V_{p,D,\lambda} := V_{d_{\mathcal{X}}^p, D, \lambda}$. The constant $L_p(g)$ of Lemma 3.2 and the constant C of condition (3) are independent of the data distribution. An exact penalty D_e therefore ensures that $V_{p,D_e,\lambda}(P, g_\sharp P_Z) = W_p^p(P, g_\sharp P_Z)$ for all $\lambda > \lambda^*(g)$ simultaneously across all distributions $P \in \mathcal{P}(\mathcal{X})$, including $\hat{P}_X^n$. This uniform exactness, in turn, enables us to derive generalization bounds uniformly over a decoder class $\mathcal{G}$.

Proposition 4.2. *Let $c = d_{\mathcal{X}}^p$ with $p \geq 1$. Suppose $\mathcal{X}$ has diameter at most K, and every decoder in $\mathcal{G}$ is γ-Lipschitz. Let the penalty D_e satisfy condition (3) with constant C. Then, for every $\lambda > \lambda_{\mathcal{G}} := pK^{p-1}\max\{1, \gamma\}/C$,*

$$\sup_{g \in \mathcal{G}} \left| V_{p,D_e,\lambda}(P_X, g_\sharp P_Z) - V_{p,D_e,\lambda}(\hat{P}_X^n, g_\sharp P_Z) \right| \leq pK^{p-1} W_p(P_X, \hat{P}_X^n).$$

Consequently, for any $s > d_p^*(P_X)$ and $\delta \in (0,1)$, the left-hand side is at most $\mathcal{O}(n^{-1/s}) + pK^p \left(\log(1/\delta)/(2n)\right)^{1/(2p)}$ with probability at least $1-\delta$.

The bound is independent of the metric entropy of $\mathcal{G}$, as the triangle inequality for W_p holds uniformly in g. The generalization gap is thus controlled solely by the intrinsic dimension of P_X. In contrast, the analogous bounds of Husain et al. (2019) require a threshold that diverges to infinity for D_f with f differentiable at 1, making the guarantee vacuous in that regime.

4.2 Rate of Convergence for Learning Distributions

Set $c = d_{\mathcal{X}}$, and let D_e, $\mathcal{G}$, and $\lambda > \lambda_{\mathcal{G}}$ be defined as in Proposition 4.2. The empirical WAE estimator is defined as $\hat{g}_n \in \arg\min_{g\in\mathcal{G}} V_{1,D_e,\lambda}(\hat{P}_X^n, g_\sharp P_Z)$. Since $V_{1,D_e,\lambda}$ coincides with W_1 on $\mathcal{G}$, the estimation error decomposes into an approximation term and a statistical term (Liang, 2021).

Proposition 4.3 (Oracle inequality)**.** *Under the assumptions of Proposition 4.2, the estimator $\hat{g}_n$ satisfies*

$$W_1(P_X, (\hat{g}_n)_\sharp P_Z) \le \inf_{g\in\mathcal{G}} W_1(P_X, g_\sharp P_Z) + 2W_1(P_X, \hat{P}_X^n).$$

If $P_X = g_\sharp^\star P_Z$ for a Lipschitz $g^\star \in \mathcal{G}$ and P_Z is supported on a compact d-dimensional differentiable manifold with $d > 2$, then $d_1^*(P_X) \le d_1^*(P_Z) \le d$ by Proposition 8 of Weed & Bach (2019) and the Lipschitz continuity of $g^\star$. It follows that $\mathbb{E}[W_1(P_X, (\hat{g}_n)_\sharp P_Z)] = \mathcal{O}(n^{-1/s})$ for every $s > d$, achieving the rate $n^{-1/d}$ up to an arbitrarily small loss in the exponent. Under model misspecification, the approximation term is generally non-vanishing; however, for Hölder $g^\star$ and sufficiently expressive neural decoders, it decays faster than the statistical term (Stéphanovitch et al., 2024).

Comparison with prior work. Prior work either suffers from a latent tolerance t with rate $\mathcal{O}(n^{-1/s} + t)$ for every $s > d$ (Chakrabarty & Das, 2021; Chakrabarty et al., 2025), or bound the excess risk of the penalized surrogate at the slower rate $\mathcal{O}(n^{-1/(d+2)})$ (Chakraborty & Bartlett, 2024). Exactness eliminates the tolerance, and directly bounds the W_1 error.

5 Experiments

The simulations of Section 5.1 fix the decoder and test whether the marginal-penalized reconstruction problem recovers a Wasserstein distance known in closed form, which restricts them to one and two dimensions. The CelebA experiment of Section 5.2 trains the decoder and examines the effect of penalty choice on latent distribution matching and local image similarity.

Optimizing the square root penalty. To compute $\sqrt{D_f}$ using unbiased stochastic gradient descent, we employed a variational form:

$$\sqrt{D_f(Q_Z, P_Z)} = \inf_{t\in\mathbb{R}} \left\{ e^{-t} D_f(Q_Z, P_Z) + e^t/4 \right\}. \tag{7}$$

We optimize the scalar t jointly with the encoder. We estimate D_f using f-GAN on the latent space (Nowozin et al., 2016), so the implemented objective approximates, rather than evaluates exactly, the population penalty; see Appendix A.

5.1 Computing a Known Wasserstein Distance

In the 1D and 2D simulations, P_X is a truncated normal distribution and P_G is constructed so that $W_2^2(P_X, P_G)$ is computable from the quantile formula (1D) or an OT map (2D). The decoder is fixed, and because the OT map is unique and deterministic, we use deterministic encoders $Q_{Z|X=x} = \delta_{q(x)}$. In 1D, the fixed coefficient is chosen from the analytically computed sufficient thresholds in Corollary 3.8. In 2D, where the required global constants are unavailable, sample estimates set an initial plug-in scale; this does not serve as an exactness certificate. Annealing increases the coefficient during training. Since the penalty never vanishes under stochastic training, we evaluate feasibility using statistical equivalence tests (Wellek, 2010) and report the relative error between reconstruction and the ground truth. Details are in Appendices D.1 and D.2.

Table 2: Simulation results with various penalties. Epochs means the number of epochs until the equivalence test was passed (at most 1000 in 1D and 500 in 2D). Rel. err. is the relative error (%) between reconstruction error and $W_2^2(P_X, P_G)$. Boldface marks the smaller absolute error within each pair $(D_f, \sqrt{D_f})$.

Dim.	Penalty	Epochs		Rel. err. (%)	
		Fixed	Anneal	Fixed	Anneal
1D	JS	1000	988 (± 24)	-2.32 (± .107)	-0.527 (± .325)
	$\sqrt{\text{JS}}$	710 (± 311)	830 (± 274)	-0.251 (± .425)	**-0.104** (± .420)
	KL	1000	842 (± 248)	-1.78 (± .100)	-0.340 (± .152)
	$\sqrt{\text{KL}}$	812 (± 266)	682 (± 310)	**-0.158** (± .223)	-0.192 (± .231)
	W_1	1000	1000	+0.148 (± .252)	+0.015 (± .196)
	MMD^2	1000	1000	-1.10 (± .039)	+0.420 (± .054)
2D	JS	428 (± 90)	358 (± 129)	-2.29 (± .099)	-0.319 (± .142)
	$\sqrt{\text{JS}}$	336 (± 164)	400 (± 97)	**-0.093** (± .160)	-0.168 (± .223)
	W_1	340 (± 151)	396 (± 148)	-0.073 (± .258)	-0.591 (± 1.04)
	MMD^2	500	400 (± 123)	-6.60 (± .060)	-0.717 (± .158)

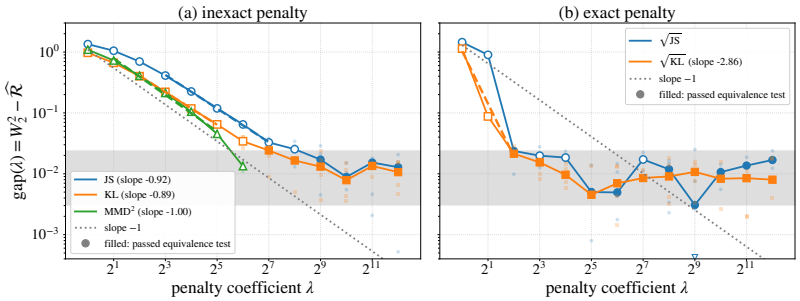


Figure 2: Inexactness gap vs. penalty coefficient in the 1D simulation (log-log scale). The gap is the shortfall of the trained reconstruction against the ground truth. Solid markers denote the median over seeds, filled upon passing the equivalence test. The shaded band marks the numerical resolution limit.

Table 2 (full results in Table S4) compares reconstruction error with the ground truth. Consistent with Proposition 3.4, the fixed-λ D_f and MMD^2 runs produce reconstruction estimates below the ground truth, and, in the 1D setting, they do not meet the equivalence tolerance within the epoch budget. The $\sqrt{D_f}$ penalties generally pass the test earlier and recover the ground truth with a relative error of at most 0.251%. Annealing λ reduces the gap for several D_f penalties, consistent with Theorem 3.3. The W_1 penalty, estimated with WGAN-LP (Petzka et al., 2018), achieves competitive accuracy but fails to pass the equivalence test, which we attribute to the numerical instability of the Lipschitz critic.

Inverse-coefficient gap. In the 1D setting, Figure 2 confirms that the empirical gaps for JS and KL trace lines with slopes close to -1 as λ grows, consistent with the $\Theta(1/\lambda)$ result. MMD^2 also underestimates the cost, but Theorem 3.5 proves only an $\Omega(1/\lambda)$ lower bound; its empirical curve is not evidence of a matching upper bound. The square-root penalties instead depart sharply from the inverse-coefficient trend and reach the numerical resolution limit at small coefficients.

5.2 Latent Distribution Matching on CelebA

We jointly train deterministic encoder and decoder on CelebA (Liu et al., 2015) with $\lambda = 100$ and logit-normal or standard normal priors (the latter covered by Section 3.4). FID (Heusel et al., 2017) evaluates generation quality. Density, Coverage (Naeem et al., 2020), Precision, and Recall (Kynkäänniemi et al., 2019) characterize the alignment between Q_Z and P_Z by measuring both prior mode coverage and latent over-concentration. These metrics are computed using 5 nearest neighbors, where Density and Coverage attain theoretical reference values of 1.0 and 0.969 when $Q_Z = P_Z$; see Appendix D.3 for further details.

Under the logit-normal prior, D_f over-concentrates latent codes, as evidenced by the shift toward smaller k-NN radii (Figure 3, left) and Density > 1. As a diagnostic of local image redundancy, we separately select the densest 1% of neighborhoods for each model and compare their decoded

Table 3: CelebA latent matching. Reference values of Density and Coverage when $Q_Z = P_Z$ are in parentheses, and boldface marks the value closer to the reference within each $(D_f, \sqrt{D_f})$ pair.

Prior	Penalty	FID ↓	Density (1.0)	Coverage (0.969)	Precision	Recall
Logit Normal	JS	57.88 (± .21)	2.365 (± .109)	0.9975 (± .0005)	0.923 (± .003)	0.407 (± .014)
	$\sqrt{\text{JS}}$	57.44 (± .23)	**0.973** (± .043)	**0.9607** (± .0055)	0.864 (± .004)	0.712 (± .012)
	KL	57.52 (± .37)	1.541 (± .043)	0.9883 (± .0010)	0.907 (± .004)	0.565 (± .009)
	$\sqrt{\text{KL}}$	57.19 (± .36)	**1.058** (± .138)	**0.9674** (± .0105)	0.851 (± .001)	0.702 (± .020)
	χ^2	57.12 (± .25)	1.555 (± .074)	0.9884 (± .0019)	0.904 (± .009)	0.554 (± .011)
	$\sqrt{\chi^2}$	57.46 (± .13)	**0.984** (± .024)	**0.9615** (± .0038)	0.864 (± .004)	0.705 (± .007)
Normal	KL	57.36 (± .45)	**1.020** (± .033)	0.9658 (± .0005)	0.697 (± .006)	0.618 (± .008)
	$\sqrt{\text{KL}}$	56.67 (± .31)	1.103 (± .025)	**0.9717** (± .0023)	0.700 (± .007)	0.611 (± .013)
	χ^2	57.39 (± .14)	**1.039** (± .028)	0.9672 (± .0019)	0.693 (± .002)	0.618 (± .008)
	$\sqrt{\chi^2}$	57.10 (± .41)	1.069 (± .046)	**0.9691** (± .0042)	0.690 (± .011)	0.616 (± .012)

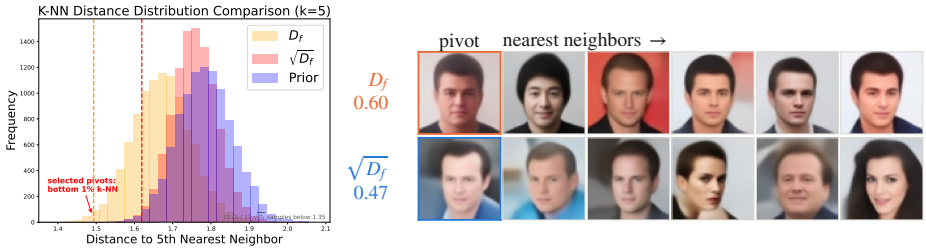


Figure 3: Logit-normal JS diagnostic. (Left) encoded k-NN distances. (Right) Decoded images of a pivot and its 5 nearest neighbors from each model's separately selected densest 1% of latent neighborhoods. Median of intra-cluster MS-SSIM values reported below each penalty.

images (Figure 3, right). In this diagnostic, $\sqrt{D_f}$ yields lower intra-cluster MS-SSIM (Wang et al., 2003) than D_f (0.513 ± 0.148 versus 0.621 ± 0.112; higher indicates similarity); see Figure S6 for the full distribution. As shown in Table 3, the $\sqrt{D_f}$ shifts Density and Coverage toward their reference values, and substantially improves Recall under the logit-normal prior; Precision decreases concurrently. FID remains comparable across penalty classes. Under the standard normal prior, the results are mixed: D_f yields better Density, whereas $\sqrt{D_f}$ yields higher Coverage.

Choice of the coefficient. We fix $\lambda = 100$ by the ablation in Table S5, the minimal tested value bringing Density and Coverage near their reference targets. The threshold in Corollary 3.8 is a conservative envelope establishing the existence of a finite exact coefficient, not a practical calibration rule for neural decoders. Our experiments thus compare various penalties at a calibrated coefficient rather than claim exactness certification; see Appendix D.4 for details.

6 Conclusion

We establish a finite-coefficient dichotomy for marginal penalties in the WAE-induced primal OT formulation. With a fixed decoder and active constraint, f-divergences with f differentiable at 1, as well as characteristic MMD^2, remain strictly below the OT cost at every finite coefficient; the gap is $\Theta(1/\lambda)$ for f satisfying generalized Pinsker condition. Conversely, the square roots of such divergences and W_1 recover the OT cost exactly above explicit thresholds. Uniform exactness over decoder classes then yields statistical guarantees for the population WAE objective.

Developing coefficient-selection algorithms remains future work. Steering methods for exact penalties (Byrd et al., 2008; Nocedal & Wright, 2006), which balance progress toward feasibility and optimality, provide a possible starting point. Our guarantees concern unrestricted encoders, population divergences, and global minimizers; restricted neural encoders, finite critics, and stochastic optimization error remain outside their scope.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the refinement and proof writing of the inexactness theorem, as well as implementing the experimental pipeline and running experiments on the CelebA dataset. We have not used generative AI tools for discovering the core research topics, formulating the main conceptual framework, proving the primary exactness results, or conducting the simulation studies; all initial drafts of the manuscript were written entirely by the authors, and other tasks subject to required disclosure are not applicable to this work. Additionally, we used generative AI tools for suggesting the paper's organization, creating and modifying scientific figures, and editing the text to improve clarity and readability. We have reviewed all AI-assisted work: all mathematical derivations and theoretical refinements assisted by LLMs were manually verified and checked for correctness by all authors, and LLM-assisted experimental code was thoroughly tested and executed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of this work, the complete source code containing implementations and instructions to reproduce all empirical experiments (including the simulation studies and CelebA experiments) is submitted as supplementary materials. Furthermore, complete proofs of all theoretical claims and detailed descriptions of all underlying assumptions and experimental setups are provided in the Appendix.

REFERENCES

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

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

Nazar Buzun, Maksim Bobrin, and Dmitry V. Dylov. Expectile regularization for fast and accurate training of neural optimal transport. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Richard H Byrd, Jorge Nocedal, and Richard A Waltz. Steering exact penalty methods for nonlinear programming. *Optimization Methods and Software*, 23(2):197–213, 2008.

Anish Chakrabarty and Swagatam Das. Statistical regeneration guarantees of the Wasserstein autoencoder with latent space consistency. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Anish Chakrabarty, Arkaprabha Basu, and Swagatam Das. Information preservation with Wasserstein autoencoders: generation consistency and adversarial robustness. *Statistics and Computing*, 35(5): 114, 2025.

Saptarshi Chakraborty and Peter L Bartlett. A statistical analysis of Wasserstein autoencoders for intrinsically low-dimensional data. In *International Conference on Learning Representations*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2016.

M. V. Dolgopolik. A unifying theory of exactness of linear penalty functions. *Optimization*, 65(6): 1167–1202, 2016.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Gustavo L Gilardoni. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Transactions on Information Theory*, 56(11):5377–5386, 2010.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hisham Husain, Richard Nock, and Robert C Williamson. A primal-dual link between GANs and autoencoders. *Advances in Neural Information Processing Systems*, 32, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? A continuous Wasserstein-2 benchmark. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alexander Korotin, Alexander Kolesov, and Evgeny Burnaev. Kantorovich strikes back! Wasserstein GANs are not optimal transport? In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *International Conference on Learning Representations*, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 2nd edition, 2006. ISBN 978-0-387-30303-1. doi: 10.1007/978-0-387-40065-5.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Giorgio Patrini, Rianne Van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pp. 733–743. PMLR, 2020.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of Wasserstein GANs. In *International Conference on Learning Representations*, 2018.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge University Press, 2025.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Wasserstein generative adversarial networks are minimax optimal distribution estimators. *The Annals of Statistics*, 52(5):2167–2193, 2024.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, volume 2, pp. 1398–1402, 2003.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019.

Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, 2010.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Computing high-dimensional optimal transport by flow neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 2872–2880. PMLR, 2025.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

A Variational Representation of f-divergences and f-GAN

An f-divergence allows for a variational representation:

$$D_f(P,Q) = \sup_{d:\mathcal{X}\to\operatorname{dom} f^*} \{\mathbb{E}_P[d] - \mathbb{E}_Q[f^*\circ d]\},$$

where $f^*(x) = \sup_y\{xy - f(y)\}$ is the convex conjugate of f and $\operatorname{dom} f^* = \{x \in \mathbb{R} : f^*(x) < \infty\}$. From this representation, it follows that D_f is convex and weakly lower semicontinuous.

Based on this representation, the f-GAN (Nowozin et al., 2016) solves

$$\min_{g\in\mathcal{G}} D_f(P_X, g_\sharp P_Z) \approx \min_{g\in\mathcal{G}} \max_{d\in\mathcal{N}} \{\mathbb{E}_{X\sim P_X}[\psi_f(d(X))] - \mathbb{E}_{Z\sim P_Z}[f^*(\psi_f(d(g(Z))))]\}$$

for a class of neural networks $\mathcal{N}$. Here, $d : \mathcal{X} \to \operatorname{dom} f^*$ is referred to as the critic or discriminator, $\psi_f : \mathbb{R} \to \operatorname{dom} f^*$ is an output activation function tailored to the specific f-divergence. It is applied to the unconstrained critic output, so that the composition $\psi_f \circ d$ takes values in $\operatorname{dom} f^*$ and plays the role of the maximizing function in the variational representation above. We leverage this variational formulation to estimate the f-divergence marginal penalty $D_f(Q_Z, P_Z)$ in (P).

B Omitted Proofs

The proofs are given in the order in which the results appear.

B.1 Proof of Corollary 3.1

We first establish the existence of a minimizer for the constrained problem (O). Given that the cost function is lower semicontinuous and Ω is a set of joint probability measures with fixed marginals P_Z and P_X, a solution to the problem $\min_{\mu\in\Omega} \mathcal{R}(\mu)$ exists. Let $\mu^* \in \Omega$ denote such a solution, so that $\mathcal{R}(\mu^*) = \mathcal{R}^*$.

Next, we consider the penalized objective. From the assumptions, it follows that for all $\nu \in A \setminus \Omega$

$$\frac{\mathcal{R}^* - \mathcal{R}(\nu)}{\mathcal{D}(\nu)} \leq \frac{L \cdot W_1(\nu, \Omega)}{C \cdot W_1(\nu, \Omega)} = \frac{L}{C}.$$

For any fixed $\lambda > L/C$, we have

$$\mathcal{R}(\nu) + \lambda\mathcal{D}(\nu) > \mathcal{R}(\nu) + \frac{L}{C}\mathcal{D}(\nu) \geq \mathcal{R}^*, \quad \forall \nu \in A \setminus \Omega.$$

Furthermore, for any $\mu \in \Omega$, we have $\mathcal{R}(\mu) + \lambda\mathcal{D}(\mu) = \mathcal{R}(\mu) \geq \mathcal{R}^*$ since $\mathcal{D}(\mu) = 0$. Combining these results, we conclude that

$$\min_{\nu\in A} \{\mathcal{R}(\nu) + \lambda\mathcal{D}(\nu)\} = \mathcal{R}^*, \quad \arg\min_{\nu\in A} \{\mathcal{R}(\nu) + \lambda\mathcal{D}(\nu)\} = \arg\min_{\mu\in\Omega} \mathcal{R}(\mu).$$

The latter equality follows from the fact that the penalized objective is strictly greater than $\mathcal{R}^*$ outside the feasible set Ω.

Remark B.1 (Solvability of the penalized problem). The exactness property inherently implies the existence of a minimizer for the penalized problem whenever the penalty coefficient exceeds the threshold, since the constrained optimal solution naturally serves as the minimizer for the penalized problem as well.

B.2 Proof of Lemma 3.2

We begin by proving a fundamental property that connects the Lipschitz continuity of h with respect to the W_1-distance to Ω.

Lemma B.2. *Let $\mathcal{R}^*$ be the global minimum of the problem (O′). If h is L-Lipschitz with respect to the W_1 metric, then $\mathcal{R}(\nu) \geq \mathcal{R}^* - L \cdot W_1(\nu, \Omega)$ for all $\nu \in A \setminus \Omega$.*

Proof of Lemma B.2. For any $\nu \in A \setminus \Omega$, consider a sequence $\{\nu_k\} \subset \Omega$ such that $W_1(\mu, \nu_k) \downarrow W_1(\mu, \Omega)$ as $k \to \infty$. By the definition of $\mathcal{R}^*$, we have $h(\nu_k) \geq \mathcal{R}^*$ for all k. Since h is L-Lipschitz,

$$\mathcal{R}^* - \mathcal{R}(\nu) \leq \mathcal{R}(\nu_k) - \mathcal{R}(\nu) \leq L \cdot W_1(\nu_k, \nu).$$

Taking the limit as $k \to \infty$ yields $\mathcal{R}^* - \mathcal{R}(\nu) \leq L \cdot W_1(\nu, \Omega)$.

□

Now it suffices to show that $\mathcal{R}$ is Lipschitz. The latter holds if the cost function $c(x, g(z)) = d_{\mathcal{X}}(x, g(z))^p$ is Lipschitz, in which case from the dual representation of the 1-Wasserstein distance we have

$$|\mathcal{R}(\mu) - \mathcal{R}(\nu)| \leq \left| \int_{\mathcal{Z}\times\mathcal{X}} c(x, g(z)) d(\mu - \nu)(z, x) \right| \leq W_1(\mu, \nu).$$

Assume g is γ-Lipschitz and $\operatorname{diam}(\mathcal{X}) \leq K$ where $\operatorname{diam}(\mathcal{X})$ is the diameter of $\mathcal{X}$. For $p \geq 1$, the function $t \mapsto t^p$ is Lipschitz on $[0, K]$ with constant pK^{p-1}. Then, for any $(z_1, x_1), (z_2, x_2) \in \mathcal{Z}\times\mathcal{X}$,

$$\begin{aligned}
|d_{\mathcal{X}}(x_1, g(z_1))^p - d_{\mathcal{X}}(x_2, g(z_2))^p| &\leq pK^{p-1}|d_{\mathcal{X}}(x_1, g(z_1)) - d_{\mathcal{X}}(x_2, g(z_2))| \\
&\leq pK^{p-1}|d_{\mathcal{X}}(x_1, x_2) + d_{\mathcal{X}}(g(z_1), g(z_2))| \\
&\leq pK^{p-1}|d_{\mathcal{X}}(x_1, x_2) + \gamma d_{\mathcal{Z}}(z_1, z_2)| \\
&\leq L_p(g) \cdot d_{\mathcal{Z}\times\mathcal{X}}((z_1^\top, x_1^\top)^\top, (z_2^\top, x_2^\top)^\top),
\end{aligned}$$

where $L_p(g) = pK^{p-1} \max\{1, \gamma\}$. Since $(z, x) \mapsto d_{\mathcal{X}}(x, g(z))^p / L_p(g)$ is 1-Lipschitz, the dual representation of W_1 gives

$$|\mathcal{R}(\mu) - \mathcal{R}(\nu)| = \left| \int_{\mathcal{Z}\times\mathcal{X}} d_{\mathcal{X}}(x, g(z))^p \, d(\mu - \nu)(z, x) \right| \leq L_p(g) \cdot W_1(\mu, \nu).$$

Hence $\mathcal{R}$ is L-Lipschitz, and Lemma B.2 yields Lemma 3.2. For $p = 1$, the same argument gives $L_1(g) = \max\{1, \gamma\}$ without any bound on $\operatorname{diam}(\mathcal{X})$.

B.3 Proof of (4)

Under the product metric on $\mathcal{Z} \times \mathcal{X}$ given by the sum of the marginal metrics, it holds that for every $\nu \in A$,

$$W_1(\pi_{1\sharp}\nu, P_Z) = W_1(\nu, \Omega). \tag{S8}$$

We establish this identity by proving both inequalities.

To this end, fix an arbitrary coupling $\gamma \in \Pi(\pi_{1\sharp}\nu, P_Z)$. We then construct a joint distribution $\nu' \in \Omega$ by defining a triplet of random variables (Z, X, Z') such that $(Z, Z') \sim \gamma$, while the conditional distribution of X given Z is identical to that of ν. Defining ν' as the joint distribution of (Z', X), we have $\pi_{1\sharp}\nu' = P_Z$ and $\pi_{2\sharp}\nu' = \pi_{2\sharp}\nu = P_X$, which ensures that $\nu' \in \Omega$.

Denote by γ' a joint distribution of (Z, X, Z', X). Then $\gamma' \in \Pi(\nu, \nu')$, and we have

$$\begin{aligned}
\mathbb{E}_{(Z,Z')\sim\gamma} d_{\mathcal{Z}}(Z, Z') &= \mathbb{E}_{(Z,X,Z',X)\sim\gamma'} d_{\mathcal{Z}}(Z, Z') \\
&= \mathbb{E}_{(Z,X,Z',X)\sim\gamma'} \left[d_{\mathcal{Z}}(Z, Z') + d_{\mathcal{X}}(X, X) \right] \\
&\geq W_1(\nu, \nu') \\
&\geq W_1(\nu, \Omega).
\end{aligned}$$

Taking the infimum over all $\gamma \in \Pi(\pi_{1\sharp}\nu, P_Z)$ yields $W_1(\pi_{1\sharp}\nu, P_Z) \geq W_1(\nu, \Omega)$.

For the reverse inequality, fix $\omega \in \Omega$ and an arbitrary coupling $\pi \in \Pi(\nu, \omega)$. Under π, the pair (Z, Z') defines a valid coupling between $\pi_{1\sharp}\nu$ and $\pi_{1\sharp}\omega = P_Z$. The transport cost $\mathbb{E}_\pi d_{\mathcal{Z}}(Z, Z')$ is bounded above by $\mathbb{E}_\pi[d_{\mathcal{Z}}(Z, Z') + d_{\mathcal{X}}(X, X')]$, which corresponds to the cost of π under the sum metric. Consequently, $W_1(\pi_{1\sharp}\nu, P_Z) \leq W_1(\nu, \omega)$ for every $\omega \in \Omega$. Taking the infimum over all $\omega \in \Omega$ yields $W_1(\pi_{1\sharp}\nu, P_Z) \leq W_1(\nu, \Omega)$, completing the proof of (S8).

B.4 Proof of Theorem 3.3

Fix $\omega \in \Omega$. Since μ_k minimizes (P′) at λ_k and $\mathcal{D}(\omega) = 0$,

$$\mathcal{R}(\mu_k) + \lambda_k \mathcal{D}(\mu_k) \leq \mathcal{R}(\omega), \qquad \mathcal{D}(\mu_k) \leq \frac{\mathcal{R}(\omega) - \mathcal{R}(\mu_k)}{\lambda_k}. \tag{S9}$$

Since $\mathcal{Z} \times \mathcal{X}$ is compact, $\mathcal{P}(\mathcal{Z} \times \mathcal{X})$ is compact under the weak topology. Therefore, there exists a subsequence $\{\mu_{k_l}\}$ that converges weakly to some measure $\mu^* \in \mathcal{P}(\mathcal{Z} \times \mathcal{X})$. The coordinate projections π_1 and π_2 are continuous, so by the continuous mapping theorem, $\pi_{2\sharp}\mu^* = P_X$, confirming that $\mu^* \in A$, and $\pi_{1\sharp}\mu_{k_l}$ weakly converges to $\pi_{1\sharp}\mu^*$.

To establish feasibility, the cost $(z, x) \mapsto c(x, g(z))$ is continuous on the compact space $\mathcal{Z} \times \mathcal{X}$ and therefore bounded. Consequently, the numerator in the right-hand side of (S9) remains bounded, so the ratio tends to 0 as $\lambda_{k_l} \uparrow \infty$. The weak lower semicontinuity of D then yields

$$\mathcal{D}(\mu^*) = D(\pi_{1\sharp}\mu^*, P_Z) \leq \liminf_{l\to\infty} D(\pi_{1\sharp}\mu_{k_l}, P_Z) \leq \liminf_{l\to\infty} \frac{\mathcal{R}(\omega) - \mathcal{R}(\mu_{k_l})}{\lambda_{k_l}} = 0.$$

Because D vanishes only on the diagonal, we obtain $\pi_{1\sharp}\mu^* = P_Z$, which proves that $\mu^* \in \Omega$.

To establish optimality, observe that the continuity and boundedness of the cost function also ensure that $\mathcal{R}(\mu_{k_l}) \to \mathcal{R}(\mu^*)$. Because $\mu^* \in \Omega$, we have $\mathcal{R}^* \leq \mathcal{R}(\mu^*)$ by definition of $\mathcal{R}^*$. Using $\mathcal{D} \geq 0$, we obtain

$$\mathcal{R}^* \leq \mathcal{R}(\mu^*) = \lim_{l\to\infty} \mathcal{R}(\mu_{k_l}) \leq \liminf_{l\to\infty} \{\mathcal{R}(\mu_{k_l}) + \lambda_{k_l} \mathcal{D}(\mu_{k_l})\} \leq \mathcal{R}(\omega).$$

Taking the infimum over all $\omega \in \Omega$ on the right-hand side gives $\mathcal{R}(\mu^*) \leq \mathcal{R}^*$, and hence $\mathcal{R}(\mu^*) = \mathcal{R}^*$. Thus, μ^* is a feasible global minimizer solving (O′). □

Remark B.3. The μ_k need only be approximate minimizers. If $\mathcal{R}(\mu_k) + \lambda_k \mathcal{D}(\mu_k) < V_{\lambda_k} + \varepsilon_k$ with $\varepsilon_k \downarrow 0$, the bound (S9) acquires an additive ε_k and the same argument applies.

B.5 An Encoder with Low Reconstruction Cost and Bounded Density

Throughout this subsection and the subsequent three, $\mathcal{X}$ and $\mathcal{Z}$ are bounded metric spaces with $\mathrm{supp}(P_Z) = \mathcal{Z}$, and the mapping $(x, z) \mapsto c(x, g(z))$ is continuous and bounded. We write

$W_c := W_c(P_X, g_\sharp P_Z)$, $c_*(x) := \inf_{z\in\mathcal{Z}} c(x, g(z))$, which is upper semicontinuous and hence Borel-measurable as an infimum of continuous functions of x, $\Delta := \mathbb{E}_{P_X}[c_*(X)] - W_c$ as defined in (5), and $V_\lambda := \inf_{\mu\in A}\{\mathcal{R}(\mu) + \lambda\mathcal{D}(\mu)\}$ for the value of (P′). Since $\mathcal{X}$ and $\mathcal{Z}$ are bounded, every conditional distribution induces a joint distribution in A, so this infimum coincides with that over $\mathcal{Q}^*$ in (P). The feasible set of (O) is nonempty because it contains the trivial encoder $Q_{Z|X=x} \equiv P_Z$; hence $W_c < \infty$.

A canonical perturbation maps each x to an optimal code $T(x) \in \arg\min_z c(x, g(z))$. However, its aggregate $T_\sharp P_X$ may be singular with respect to P_Z. For f-divergences satisfying $f'(\infty) = \infty$, such as KL and χ^2, the divergence along the mixture path would then be infinite for every $\alpha > 0$. For divergences with $f'(\infty) < \infty$, such as JS, the singular component would contribute a term linear in α. We therefore restrict the prior to sublevel sets of minimal reconstruction cost rather than concentrating mass on a single minimizer.

For $\theta \in (0, 1)$ and $x \in \mathcal{X}$, let $F(x, s) := P_Z(\{z : c(x, g(z)) \le s\})$, $t_\theta(x) := \inf\{s \ge 0 : F(x, s) \ge \theta\}$, $A_\theta(x) := \{z : c(x, g(z)) \le t_\theta(x)\}$, and $\theta(x) := P_Z(A_\theta(x))$. The right-continuity of $s \mapsto F(x, s)$ implies $\theta(x) \ge \theta$. The *quantile encoder* is defined by

$$R^\theta_{Z|X=x}(dz) := \frac{\mathbf{1}\{z \in A_\theta(x)\}}{\theta(x)}\, P_Z(dz),$$

representing the normalized conditional prior restricted to the θ-sublevel set of minimal reconstruction cost for the observation x.

Lemma B.4 (Measurability and bounded density)**.** *The maps $x \mapsto t_\theta(x)$ and $x \mapsto \theta(x)$ are Borel-measurable, the set $\{(x, z) : z \in A_\theta(x)\}$ is jointly measurable, and the aggregate R^θ_Z is absolutely continuous with respect to P_Z with density $\phi_\theta \le 1/\theta$, P_Z-almost everywhere. In particular, $\chi^2(R^\theta_Z, P_Z) \le 1/\theta - 1$.*

Proof. For fixed s, the map $x \mapsto F(x, s) = \int \mathbf{1}\{c(x, g(z)) \le s\}\, dP_Z(z)$ is Borel-measurable by Tonelli's theorem, since $(x, z) \mapsto c(x, g(z))$ is continuous. Because $F(x, \cdot)$ is nondecreasing and right-continuous, we have $\{x : t_\theta(x) \le s\} = \{x : F(x, s) \ge \theta\}$, establishing that t_θ is Borel. The set $\{(x, z) : c(x, g(z)) - t_\theta(x) \le 0\}$ is therefore jointly measurable, and $\theta(x) = F(x, t_\theta(x))$ is Borel. For a Borel set $B \subset \mathcal{Z}$, Tonelli's theorem yields $R^\theta_Z(B) = \int_B \phi_\theta\, dP_Z$, where

$$\phi_\theta(z) := \int \frac{\mathbf{1}\{z \in A_\theta(x)\}}{\theta(x)}\, dP_X(x) \le \frac{1}{\theta},$$

since $\theta(x) \ge \theta$. Finally, $\int \phi_\theta^2\, dP_Z \le \theta^{-1} \int \phi_\theta\, dP_Z = \theta^{-1}$. □

Lemma B.5 (Cost of the quantile encoder)**.** *The function $G(\theta) := \mathbb{E}_{R^\theta_{Z|X} P_X}\, c(X, g(Z))$ is nondecreasing on $(0, 1)$, and $G(\theta) \to W_c + \Delta$ as $\theta \downarrow 0$.*

Proof. Monotonicity. Fix x and choose $\theta_1 < \theta_2$, so that $A_{\theta_1}(x) \subseteq A_{\theta_2}(x)$. Defining $\bar{c}_1 := \theta_1(x)^{-1} \int_{A_{\theta_1}(x)} c(x, g(z))\, dP_Z(z)$, it follows that $\bar{c}_1 \le t_{\theta_1}(x)$, and $c(x, g(z)) > t_{\theta_1}(x) \ge \bar{c}_1$ on $A_{\theta_2}(x) \setminus A_{\theta_1}(x)$. Hence

$$\frac{1}{\theta_2(x)} \int_{A_{\theta_2}(x)} c(x, g(z))\, dP_Z(z) \ge \frac{\theta_1(x)}{\theta_2(x)}\bar{c}_1 + \frac{\theta_2(x) - \theta_1(x)}{\theta_2(x)}\bar{c}_1 = \bar{c}_1,$$

and integrating over P_X gives $G(\theta_1) \le G(\theta_2)$.

Limit. Fix $\eta > 0$. By the continuity of c and g, the set $U_\eta(x) := \{z : c(x, g(z)) < c_*(x) + \eta\}$ is open and nonempty, and therefore $P_Z(U_\eta(x)) > 0$ due to the full support assumption $\mathrm{supp}(P_Z) = \mathcal{Z}$. For $\theta < P_Z(U_\eta(x))$ we have $F(x, c_*(x) + \eta) \ge P_Z(U_\eta(x)) > \theta$, hence $t_\theta(x) \le c_*(x) + \eta$ and $\mathbb{E}_{R^\theta_{Z|X=x}}\, c(x, g(Z)) \le c_*(x) + \eta$. Thus $\mathbb{E}_{R^\theta_{Z|X=x}}\, c(x, g(Z)) \to c_*(x)$ pointwise for every x. Since the integrands are uniformly bounded, the dominated convergence theorem yields $G(\theta) \to \mathbb{E}_{P_X}[c_*(X)] = W_c + \Delta$. □

Since $\Delta < 0$, Lemma B.5 guarantees the existence of $\theta^* \in (0, 1)$ such that $G(\theta^*) \le W_c + \Delta/2 < W_c$. Fixing such a θ^*, we denote the resulting quantile encoder by $R^*_{Z|X} := R^{\theta^*}_{Z|X}$, its aggregate marginal on $\mathcal{Z}$ by R^*_Z, its Radon–Nikodym derivative by $\phi^* := \phi_{\theta^*}$, and set $M^* := 1/\theta^* > 1$. Write $\rho = R^*_{Z|X} P_X \in A$ for the induced joint distribution, so that $\pi_{1\sharp}\rho = R^*_Z$ and $\mathcal{R}(\rho) = G(\theta^*)$.

Lemma B.6 (Nondegeneracy). $R_Z^* \neq P_Z$. *Consequently* $\Gamma := \chi^2(R_Z^*, P_Z) \in (0, M^* - 1]$, *and* $\mathrm{MMD}_k^2(R_Z^*, P_Z) > 0$ *for every characteristic kernel* k.

Proof. If $R_Z^* = P_Z$, then R would be feasible for (O), implying $G(\theta^*) \geq W_c$, which contradicts $G(\theta^*) < W_c$. The upper bound on Γ is Lemma B.4. Finally, if k is characteristic, the injectivity of the mean embedding and $R_Z^* \neq P_Z$ imply $\mathrm{MMD}_k^2(R_Z^*, P_Z) > 0$. □

The perturbation path. Fix $\varepsilon > 0$ and choose a feasible joint distribution $\bar{\mu} \in \Omega$ satisfying $\mathcal{R}(\bar{\mu}) \leq W_c + \varepsilon$. For $\alpha \in [0, 1]$, define

$$\mu_\alpha := (1 - \alpha)\bar{\mu} + \alpha\rho.$$

Its aggregate distribution is $\pi_{1\sharp}\mu_\alpha = (1 - \alpha)P_Z + \alpha R_Z^*$, with Radon–Nikodym derivative

$$\frac{d\pi_{1\sharp}\mu_\alpha}{dP_Z} = 1 + \alpha(\phi^* - 1) \in \big[1 - \alpha,\ 1 + \alpha(M^* - 1)\big] \quad P_Z\text{-a.e.}$$

Furthermore, by linearity of the reconstruction cost with respect to the encoder,

$$\mathcal{R}(\mu_\alpha) = (1 - \alpha)\mathcal{R}(\bar{\mu}) + \alpha\mathcal{R}(\rho) \leq (1 - \alpha)(W_c + \varepsilon) + \alpha\Big(W_c + \frac{\Delta}{2}\Big) \leq W_c + \varepsilon - \frac{|\Delta|}{2}\alpha. \quad \text{(S10)}$$

B.6 Proof of Proposition 3.4

The case $D = D_f$. For $\alpha \in [0, 1/2]$ the density $1 + \alpha(\phi^* - 1)$ takes values in the compact interval $I := [1/2, (1 + M^*)/2] \subset (0, \infty)$, on which the convex and finite function f is Lipschitz with some constant ℓ_I. Using $f(1) = 0$, we obtain $|f(1 + \alpha(\phi^* - 1))|/\alpha \leq \ell_I|\phi^* - 1| \leq \ell_I M^*$. Moreover, because f is differentiable at 1, $f(1 + \alpha(\phi^* - 1))/\alpha \to f'(1)(\phi^* - 1)$ pointwise as $\alpha \downarrow 0$. Since $\pi_{1\sharp}\mu_\alpha$ is absolutely continuous with respect to P_Z, dominated convergence yields

$$\frac{\mathcal{D}(\mu_\alpha)}{\alpha} = \frac{D_f(\pi_{1\sharp}\mu_\alpha, P_Z)}{\alpha} = \int \frac{f(1 + \alpha(\phi^* - 1))}{\alpha}\, dP_Z \to f'(1) \int (\phi^* - 1)\, dP_Z = 0.$$

Fix $\lambda > 0$ and select $\alpha \in (0, 1/2]$ such that $\lambda\mathcal{D}(\mu_\alpha) \leq \alpha|\Delta|/4$. This choice depends on $R_{Z|X}^*$ but is independent of $\bar{\mu}$ and ε, because $\mathcal{D}(\mu_\alpha)$ depends only on R_Z^*. By (S10), $V_\lambda \leq \mathcal{R}(\mu_\alpha) + \lambda\mathcal{D}(\mu_\alpha) \leq W_c + \varepsilon - \alpha|\Delta|/4$, and letting $\varepsilon \downarrow 0$ gives $V_\lambda \leq W_c - \alpha|\Delta|/4 < W_c$.

The case $D = \mathrm{MMD}_k^2$. Recall that $m_P := \int k(\cdot, z)\, dP(z)$ denotes the kernel mean embedding. Then $m_{\pi_{1\sharp}\mu_\alpha} - m_{P_Z} = \alpha(m_{R_Z^*} - m_{P_Z})$, which implies

$$\mathrm{MMD}_k^2(\pi_{1\sharp}\mu_\alpha, P_Z) = \alpha^2\, \mathrm{MMD}_k^2(R_Z^*, P_Z) =: \alpha^2\mathfrak{M}.$$

Boundedness of k ensures that $\mathfrak{M} < \infty$. Then $\mathcal{R}(\mu_\alpha) + \lambda\mathcal{D}(\mu_\alpha) \leq W_c + \varepsilon - \alpha|\Delta|/2 + \lambda\mathfrak{M}\alpha^2$, which is strictly bounded above by $W_c + \varepsilon - \alpha|\Delta|/4$ for sufficiently small $\alpha > 0$. Taking $\varepsilon \downarrow 0$ concludes the proof. □

B.7 Proof of Theorem 3.5

Lower bound on the gap. Define $K_0 := \frac{1}{2}\max_{t \in I} f''(t)$, which is finite because $f \in C^2(0, \infty)$ on the compact interval I, and strictly positive because $1 \in I$ and $f''(1) > 0$. Applying Taylor's theorem with Lagrange remainder at $t = 1$ and using $f(1) = 0$, we have $f(1 + \alpha(\phi^* - 1)) \leq f'(1)\alpha(\phi^* - 1) + K_0\alpha^2(\phi^* - 1)^2$ for all $\alpha \in [0, 1/2]$. Integrating against P_Z, the linear term vanishes and $\mathcal{D}(\mu_\alpha) = D_f(\pi_{1\sharp}\mu_\alpha, P_Z) \leq K_0\Gamma\alpha^2$. Combining this with (S10),

$$\mathcal{R}(\mu_\alpha) + \lambda\mathcal{D}(\mu_\alpha) \leq W_c + \varepsilon - \frac{|\Delta|}{2}\,\alpha + \lambda K_0\Gamma\alpha^2.$$

The right-hand side is minimized at $\alpha^\dagger := |\Delta|/(4\lambda K_0\Gamma)$, which lies in $[0, 1/2]$ precisely when $\lambda \geq \lambda_0 := |\Delta|/(2K_0\Gamma)$, with minimum value $W_c + \varepsilon - \Delta^2/(16\lambda K_0\Gamma)$. Letting $\varepsilon \downarrow 0$, we obtain

$$W_c - V_\lambda \geq \frac{\Delta^2}{16K_0\Gamma} \cdot \frac{1}{\lambda} \qquad \text{for all } \lambda \geq \lambda_0,$$

establishing the lower bound with $c_1 := \Delta^2/(16K_0\Gamma)$. For $0 < \lambda < \lambda_0$, setting $\alpha = 1/2$ yields $W_c - V_\lambda \geq |\Delta|/4 - \lambda K_0\Gamma/4 > |\Delta|/8$. Therefore, for all $\lambda > 0$

$$W_c - V_\lambda \geq \min\left\{\frac{c_1}{\lambda}, \frac{|\Delta|}{8}\right\}.$$

For $D = \mathrm{MMD}_k^2$ with a characteristic kernel, an analogous derivation gives

$$\mathcal{R}(\mu_\alpha) + \lambda\mathcal{D}(\mu_\alpha) \leq W_c + \varepsilon - \frac{|\Delta|}{2}\alpha + \lambda\mathfrak{M}\alpha^2, \qquad \alpha \in [0,1],$$

where $\mathfrak{M} > 0$ by Lemma B.6. If $\lambda \geq |\Delta|/(4\mathfrak{M})$, the minimizer $\alpha^\dagger = |\Delta|/(4\lambda\mathfrak{M})$ belongs to $[0,1]$ and yields $W_c - V_\lambda \geq \Delta^2/(16\mathfrak{M}\lambda)$. If $0 < \lambda < |\Delta|/(4\mathfrak{M})$, taking $\alpha = 1$ yields $W_c - V_\lambda \geq |\Delta|/2 - \lambda\mathfrak{M} > |\Delta|/4$. This proves the claimed lower bound.

Upper bound on the gap. Set $L := L_p(g)$. Let $\mu \in A$ and $u := \mathrm{TV}(\pi_{1\sharp}\mu, P_Z)$. If $\mu \in \Omega$, then $\mathcal{R}(\mu) + \lambda\mathcal{D}(\mu) \geq W_c$. Otherwise Lemma 3.2, (4), and the inequality $W_1 \leq \mathrm{diam}(\mathcal{Z})\,\mathrm{TV}$ yield

$$\mathcal{R}(\mu) \geq W_c - L\,W_1(\mu,\Omega) \geq W_c - L\,W_1(\pi_{1\sharp}\mu, P_Z) \geq W_c - L\,\mathrm{diam}(\mathcal{Z})\,u,$$

and the generalized Pinsker inequality (Proposition B.7) ensures $D_f(\pi_{1\sharp}\mu, P_Z) \geq 2f''(1)u^2$. Consequently,

$$\mathcal{R}(\mu) + \lambda\mathcal{D}(\mu) \geq W_c - L\,\mathrm{diam}(\mathcal{Z})\,u + 2\lambda f''(1)\,u^2 \geq W_c - \frac{(L\,\mathrm{diam}(\mathcal{Z}))^2}{8\lambda f''(1)},$$

where the final inequality follows by minimizing the quadratic form over $u \geq 0$. Taking the infimum over $\mu \in A$ completes the proof. □

Reconstruction gap (6) at a global minimizer. Set $L = L_p(g)$ and write $u = \mathrm{TV}(\pi_{1\sharp}\mu^\lambda, P_Z)$. Optimality and feasibility of a constrained optimizer imply

$$2\lambda f''(1)u^2 \leq \lambda D_f(\pi_{1\sharp}\mu^\lambda, P_Z) \leq W_c - \mathcal{R}(\mu^\lambda) \leq L\,\mathrm{diam}(\mathcal{Z})u.$$

If $u > 0$, then $u \leq L\,\mathrm{diam}(\mathcal{Z})/(2\lambda f''(1))$ and $W_c - \mathcal{R}(\mu^\lambda) \leq (L\,\mathrm{diam}(\mathcal{Z}))^2/(2\lambda f''(1))$; if $u = 0$, $W_c - \mathcal{R}(\mu^\lambda) = 0$. The lower bound follows from $W_c - \mathcal{R}(\mu^\lambda) \geq W_c - V_\lambda \geq c_1/\lambda$, proving (6).

B.8 Proof of Remark 3.6

Since every feasible encoder incurs a reconstruction cost of at least $\mathbb{E}_{P_X}[c_*(X)]$, it follows that $\Delta \leq 0$. Let $S := \{(z,x) : c(x, g(z)) = c_*(x)\}$, which is Borel-measurable because c_* is Borel-measurable. If the infimum is not attained for any x, then $S = \emptyset$, and the argument below directly yields $\Delta < 0$.

If some coupling $\pi \in \Pi(P_Z, P_X)$ satisfies $\pi(S) = 1$, its conditional distribution of Z given X defines a feasible encoder with cost $\mathbb{E}_{P_X}[c_*(X)]$, implying $W_c \leq \mathbb{E}_{P_X}[c_*(X)]$ and hence $\Delta = 0$. Conversely, suppose $\Delta = 0$. Because couplings with fixed marginals form a tight, weakly closed family and the cost function is lower semicontinuous, there exists an optimal coupling $\pi^* \in \Pi(P_Z, P_X)$ that attains W_c. Then $\mathbb{E}_{\pi^*}[c(X, g(Z)) - c_*(X)] = W_c - \mathbb{E}_{P_X}[c_*(X)] = 0$ with a nonnegative integrand, so $\pi^*(S) = 1$. Finally, if the minimizer $T(x)$ is unique for P_X-almost every x, any coupling concentrated on S satisfies $Z = T(X)$ almost surely and therefore admits first marginal $T_\sharp P_X$. If $T_\sharp P_X \neq P_Z$, no such coupling can belong to $\Pi(P_Z, P_X)$, which implies $\Delta < 0$. □

Where the total variation distance escapes. For the total variation penalty corresponding to a f-divergence with $f(t) = \frac{1}{2}|t-1|$, this lower-bound construction fundamentally fails. Along the perturbation path μ_α, $\mathrm{TV}(\pi_{1\sharp}\mu_\alpha, P_Z) = \alpha\,\mathrm{TV}(R_Z^*, P_Z)$ scales linear in α, so the penalty offsets the linear cost decrease of (S10) as soon as $\lambda > |\Delta|/(2\,\mathrm{TV}(R_Z^*, P_Z))$, consistent with Theorem 3.7(b). The differentiability of f at 1 is the precise mechanism ensuring that $D_f(\pi_{1\sharp}\mu_\alpha, P_Z) = o(\alpha)$ in the proof of Proposition 3.4.

B.9 Proof of Theorem 3.7

We utilize the following result to link f-divergences to the total variation distance.

Proposition B.7 (Generalized Pinsker's inequality (Gilardoni, 2010, Corollary 4))**.** *Assume that the convex function f defining D_f is three times differentiable and $f''(u) > 0$ for all $u \neq 1$. If f satisfies*

$$\operatorname{sign}(u-1)\left\{\frac{f'''(u)}{f''(u)}\left[1-\frac{c}{3}(u-1)\right]-c\right\} \geq 0, \tag{S11}$$

where $c = f'''(1)/f''(1)$, then

$$D_f(P,Q) \geq 2f''(1)\cdot TV(P,Q)^2.$$

Verification for the f-GAN divergences. Write $r(u) := f'''(u)/f''(u)$ and $c := r(1)$, so that the left-hand side of (S11) is $\operatorname{sign}(u-1)\{r(u)[1-\frac{c}{3}(u-1)]-c\}$. A direct computation gives the values below, all nonnegative for $u > 0$, so each of these divergences satisfies Proposition B.7.

	$f(u)$	$f''(1)$	c	left-hand side of (S11)
KL	$u\log u$	1	-1	$2\lvert u-1\rvert/(3u)$
reverse KL	$-\log u$	1	-2	$2\lvert u-1\rvert/(3u)$
JS	$u\log u-(u+1)\log\frac{u+1}{2}$	$1/2$	$-3/2$	$\lvert u-1\rvert/(2u)$
squared Hellinger	$(\sqrt{u}-1)^2$	$1/2$	$-3/2$	$3\lvert u-1\rvert/(4u)$
χ^2	$(u-1)^2$	2	0	0
reverse χ^2	$(u-1)^2/u$	2	-3	0

Proof of Theorem 3.7. If $\nu \in \Omega$, the LHS and RHS of (3) are both 0 and there is nothing to show. Now assume $\nu \in A \setminus \Omega$. Proposition B.7 implies that $\sqrt{D_f(\pi_{1\sharp}\nu, P_Z)} \geq \sqrt{2f''(1)}\cdot TV(\pi_{1\sharp}\nu, P_Z)$. Furthermore, since $\mathcal{Z}$ is bounded, $TV(\pi_{1\sharp}\nu, P_Z) \geq W_1(\pi_{1\sharp}\nu, P_Z)/\operatorname{diam}(\mathcal{Z})$ holds (Villani, 2009, Theorem 6.15). Combining these yields

$$\sqrt{D_f(\pi_{1\sharp}\nu, P_Z)} \geq \sqrt{2f''(1)}\cdot TV(\pi_{1\sharp}\nu, P_Z) \geq \frac{\sqrt{2f''(1)}}{\operatorname{diam}(\mathcal{Z})}\cdot W_1(\pi_{1\sharp}\nu, P_Z). \tag{S12}$$

Together with (S8), it proves part (a). Part (b) follows from $TV \geq W_1/\operatorname{diam}(\mathcal{Z})$ and (S8), and part (c) is (S8) itself.

B.10 Proof of Proposition 3.9

If ν is not absolutely continuous with respect to P_Z, then $\chi^2(\nu, P_Z) = \infty$, and the inequality holds trivially. Assume then that ν is absolutely continuous with respect to P_Z.

By the duality, the 1-Wasserstein distance is given by

$$W_1(\nu, P_Z) = \sup_{\phi\in\mathcal{H}_1}\left|\int \phi\, d(\nu - P_Z)\right|.$$

For any 1-Lipschitz function $\phi : \mathcal{Z} \to \mathbb{R}$, we have

$$\begin{aligned}\left|\int \phi\, d(\nu - P_Z)\right| &= \left|\int_{\mathcal{Z}} (\phi - \mathbb{E}_{P_Z}[\phi])\left(\frac{d\nu}{dP_Z} - 1\right) dP_Z\right| \\ &\leq \sqrt{\int (\phi - \mathbb{E}_{P_Z}[\phi])^2\, dP_Z} \times \sqrt{\int \left(\frac{d\nu}{dP_Z} - 1\right)^2 dP_Z} \\ &= \sqrt{\int (\phi - \mathbb{E}_{P_Z}[\phi])^2\, dP_Z} \times \sqrt{\chi^2(\nu, P_Z)},\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality in $L^2(P_Z)$.

To bound the variance term, let $Z_1, Z_2 \sim P_Z$ be i.i.d. random variables. We use the identity

$$\mathrm{Var}_{P_Z}(\phi(Z_1)) = \frac{1}{2}\mathbb{E}\left[(\phi(Z_1) - \phi(Z_2))^2\right].$$

Since ϕ is 1-Lipschitz, $(\phi(Z_1) - \phi(Z_2))^2 \le \|Z_1 - Z_2\|^2$. Taking the expectation on both sides and using $\mathbb{E}\|Z_1 - Z_2\|^2 = 2\,\mathrm{Var}_{P_Z}(Z) \le 2\,\mathbb{E}_{P_Z}\|Z\|^2$ yields $\mathrm{Var}_{P_Z}(\phi(Z_1)) \le \mathbb{E}_{P_Z}\|Z\|^2 =: \sigma^2$. Then we obtain

$$\left|\int \phi\, d(\nu - P_Z)\right| \le \sigma\sqrt{\chi^2(\nu, P_Z)}.$$

Taking the supremum over all 1-Lipschitz functions yields $W_1(\nu, P_Z) \le \sigma\sqrt{\chi^2(\nu, P_Z)}$. Combining this with (4) gives condition (3) with $C = \sigma^{-1}$, which concludes the proof.

B.11 Proof of Proposition 4.2

Let $L_S(g) := V_{p,D_e,\lambda}(P_X, g_\sharp P_Z) - V_{p,D_e,\lambda}(\hat{P}_X^n, g_\sharp P_Z)$. For $c = d_{\mathcal{X}}^p$, Lemma 3.2 verifies condition (2) with constant $L = pK^{p-1}\max\{1, \mathrm{Lip}(g)\} \le pK^{p-1}\max\{1, \gamma\}$ for every $g \in \mathcal{G}$, independently of the underlying distribution on $\mathcal{X}$. Moreover, condition (3) holds with the identical constant C across all data distributions, because its proof depends only on D_e, P_Z, and (4), which preserves the marginal distribution on $\mathcal{X}$. By Corollary 3.1, for $\lambda > \lambda_{\mathcal{G}}$, the penalty is therefore exact simultaneously for P_X and $\hat{P}_X^n$, yielding

$$L_S(g) = W_p^p(P_X, g_\sharp P_Z) - W_p^p(\hat{P}_X^n, g_\sharp P_Z) \qquad \text{for every } g \in \mathcal{G}.$$

Because both Wasserstein distances take values in $[0, K]$, and $|a^p - b^p| \le pK^{p-1}|a - b|$ for $a, b \in [0, K]$ by the mean value theorem, the triangle inequality for W_p implies $|L_S(g)| \le pK^{p-1}W_p(P_X, \hat{P}_X^n)$. Since the right-hand side does not depend on g, we obtain

$$\sup_{g\in\mathcal{G}} |L_S(g)| \le pK^{p-1}W_p(P_X, \hat{P}_X^n) \tag{S13}$$

for every realization of the empirical measure $\hat{P}_X^n$. Crucially, this bound holds independently of the covering numbers or metric entropy of the decoder class $\mathcal{G}$.

For the probability bound, fix $s > d_p^*(P_X)$. By Proposition 5 of Weed & Bach (2019), the expected empirical Wasserstein distance satisfies $\mathbb{E}[W_p^p(P_X, \hat{P}_X^n)] = \mathcal{O}(n^{-p/s})$. Applying the bounded difference inequality to W_p^p (Weed & Bach, 2019, Proposition 20), rescaled by the diameter bound K^p, yields

$$\mathbb{P}\left(W_p^p(P_X, \hat{P}_X^n) \ge \mathbb{E}[W_p^p(P_X, \hat{P}_X^n)] + \epsilon\right) \le \exp\left\{-\frac{2n\epsilon^2}{K^{2p}}\right\}.$$

Solving $\delta = \exp\{-2n\epsilon^2/K^{2p}\}$ for ϵ and using $(a+b)^{1/p} \le a^{1/p} + b^{1/p}$ for $a, b \ge 0$ and $p \ge 1$, with probability at least $1 - \delta$,

$$W_p(P_X, \hat{P}_X^n) \le \left(\mathcal{O}(n^{-p/s}) + K^p\sqrt{\tfrac{\log(1/\delta)}{2n}}\right)^{1/p} \le \mathcal{O}(n^{-1/s}) + K\left(\tfrac{\log(1/\delta)}{2n}\right)^{1/(2p)}.$$

Combining this deviation bound with (S13) completes the proof. In particular, because $s > d_p^*(P_X) \ge 2p$ by Definition 4.1, we have $1/s < 1/(2p)$, meaning that the non-parametric rate $\mathcal{O}(n^{-1/s})$ dominates the parametric deviation term asymptotically. □

B.12 Proof of Proposition 4.3

By triangle inequality, we have, for any $g \in \mathcal{G}$

$$\begin{aligned} W_1(P_X, (\hat{g}_n)_\sharp P_Z) &\le W_1(P_X, \hat{P}_X^n) + W_1(\hat{P}_X^n, (\hat{g}_n)_\sharp P_Z) \\ &\le W_1(P_X, \hat{P}_X^n) + W_1(\hat{P}_X^n, g_\sharp P_Z) \\ &\le 2W_1(P_X, \hat{P}_X^n) + W_1(P_X, g_\sharp P_Z). \end{aligned}$$

Taking the infimum over $\mathcal{G}$ yields the oracle inequality.

C Proof of Equation (1) within the Exact Penalty Framework

Let A_p and Ω_p denote the domain and feasible set, respectively, obtained by replacing the 1-Wasserstein space with the p-Wasserstein space in the definitions of A and Ω in problem (O′). Furthermore, consider the following constrained optimization problem

$$\min_{\mu \in A_p} [\mathbb{E}_{(Z,X)\sim\mu} \|X - g(Z)\|_p^p]^{1/p} \quad \text{s.t.} \quad \pi_{1\sharp}\mu = P_Z,$$

and its corresponding penalized formulation

$$\min_{\mu \in A_p} \left\{ [\mathbb{E}_{(Z,X)\sim\mu} \|X - g(Z)\|_p^p]^{1/p} + \lambda \cdot W_p(\pi_{1\sharp}\mu, P_Z) \right\}.$$

The two conditions of Theorem 2.1 (and Corollary 3.1) can be written as follows for the p-Wasserstein distance,

$$\begin{aligned} [\mathbb{E}_{(Z,X)\sim\mu} \|X - g(Z)\|_p^p]^{1/p} &\geq W_p(P_X, P_G) - \gamma \cdot W_p(\mu, \Omega_p), \quad \forall \mu \in A_p \setminus \Omega_p, \\ W_p(\pi_{1\sharp}\mu, P_Z) &\geq W_p(\mu, \Omega_p), \quad \forall \mu \in A_p, \end{aligned}$$

where γ denotes the Lipschitz constant of the decoder g and $W_p(\mu, \Omega_p) := \inf_{\omega \in \Omega_p} W_p(\mu, \omega)$.

The second inequality is analogous to the inequality established in (S8). Let $\nu^* \in \Pi(\pi_{1\sharp}\mu, P_Z)$ be an optimal joint distribution that achieves the p-Wasserstein distance $W_p(\pi_{1\sharp}\mu, P_Z)$. By the gluing lemma (Villani, 2009), we can construct a joint distribution $\omega \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z} \times \mathcal{X})$ for the random triplet (Z, Z', X) such that its (Z, X)-marginal is μ and its (Z, Z')-marginal is ν^*. Let τ denote the (Z', X)-marginal of ω. By construction, $\tau \in \Omega_p$.

By considering the coupling of μ and τ induced by ω, the transportation cost is determined solely by the displacement within the $\mathcal{Z}$ space, as the X components are identical. Consequently, we obtain

$$W_p(\mu, \tau) \leq [\mathbb{E}_{\nu^*} \|Z - Z'\|_p^p]^{1/p} = W_p(\pi_{1\sharp}\mu, P_Z).$$

Taking the infimum over $\tau \in \Omega_p$ yields $W_p(\mu, \Omega_p) \leq W_p(\pi_{1\sharp}\mu, P_Z)$.

To show the first inequality, applying Minkowski's inequality yields

$$\begin{aligned} [\mathbb{E}_{(Z,X)\sim\mu} \|X - g(Z)\|_p^p]^{1/p} &= [\mathbb{E}_{(Z,X,Z')\sim\omega} \|X - g(Z)\|_p^p]^{1/p} \\ &\geq [\mathbb{E}_{\omega} \|X - g(Z')\|_p^p]^{1/p} - [\mathbb{E}_{\omega} \|g(Z) - g(Z')\|_p^p]^{1/p} \\ &= [\mathbb{E}_{\tau} \|X - g(Z')\|_p^p]^{1/p} - [\mathbb{E}_{\nu^*} \|g(Z) - g(Z')\|_p^p]^{1/p} \\ &\geq W_p(P_X, P_G) - \gamma \, [\mathbb{E}_{\nu^*} \|Z - Z'\|_p^p]^{1/p} \\ &= W_p(P_X, P_G) - \gamma \cdot W_p(\pi_{1\sharp}\mu, P_Z). \end{aligned}$$

To complete the proof of the first condition, it remains to show that $W_p(\mu, \Omega_p) \geq W_p(\pi_{1\sharp}\mu, P_Z)$, which corresponds to the reverse inequality of the second condition. For any arbitrary $\omega' \in \Omega_p$, we have

$$\begin{aligned} W_p(\mu, \omega') &= \left\{ \inf_{\pi \in \Pi(\mu, \omega')} \mathbb{E}_{(Z,X,Z',X')\sim\pi} \left\| (Z, X)^\top - (Z', X')^\top \right\|_p^p \right\}^{1/p} \\ &\geq \left\{ \inf_{\pi \in \Pi(\mu, \omega')} \mathbb{E}_{(Z,X,Z',X')\sim\pi} \|Z - Z'\|_p^p \right\}^{1/p} \\ &\geq W_p(\pi_{1\sharp}\mu, \pi_{1\sharp}\omega') \\ &= W_p(\pi_{1\sharp}\mu, P_Z). \end{aligned}$$

Taking the infimum over all $\omega' \in \Omega_p$ yields $W_p(\pi_{1\sharp}\mu, P_Z) \leq W_p(\mu, \Omega_p)$. Substituting this into the above inequality completes the proof.

As a result, the sufficient exactness threshold λ^* in our marginal-penalty framework evaluates to γ (with $L_p(g) = \gamma$ and $C = 1$), which coincides with the threshold of Patrini et al. (2020).

D FURTHER DETAILS ON THE EXPERIMENTS

Computational resources. The simulation experiments were conducted on a machine equipped with an Intel Xeon Silver 4114 CPU (2.20 GHz) and an NVIDIA Titan V GPU with 12 GB of memory, using Python 3.10.12, PyTorch 2.0.1, CUDA 12.2, and cuDNN 8.9.5.

For the CelebA experiments, we utilize a high-performance server featuring an Intel Xeon Platinum 8568Y+ CPU (4.0 GHz) with an NVIDIA RTX 6000 Ada GPU (48 GB). This setup ran on Python 3.10.12, PyTorch 2.9.1, PyTorch Lightning 2.6.0, CUDA 12.2, and cuDNN 8.9.5.

Our code is provided in the supplementary material.

D.1 ONE-DIMENSIONAL SIMULATION

Setup. For the one-dimensional simulation, the P_G is set as a truncated normal distribution. The parameters and supports of distributions P_X and P_G were randomly initialized. The P_X was defined as a normal distribution $N(-1.172, 0.741^2)$ truncated to $[-4.135, 1.790]$, and P_G was defined as $N(0.526, 0.756^2)$ truncated to $[-0.229, 1.282]$. To compute the ground-truth squared 2-Wasserstein distance between these distributions, we leveraged a closed form of the squared 2-Wasserstein distance

$$W_2^2(P_X, P_G) = \int_0^1 |F_X^{-1}(s) - F_G^{-1}(s)|^2 ds, \tag{S14}$$

where F_X and F_G are the cumulative distribution functions (CDFs) of P_X and P_G, respectively. We used (S14) to numerically compute the ground-truth value of (O), $W_2^2(P_X, P_G) = 3.003$, which serve as the benchmark for evaluating the relative errors in Tables 2 and S4. The prior P_Z was set to unif$([0, 1])$, implying the decoder g acts as F_G^{-1}. In this setup, since the OT map from P_X to P_G is given by the monotone rearrangement $F_G^{-1} \circ F_X = g \circ F_X$, the optimal encoder is $Q^*_{Z|X=x} = \delta_{F_X(x)}$. Indeed, the marginal precisely matches the prior, $Q_Z = (F_X)_\sharp P_X = P_Z$. Consequently, it is sufficient to restrict the encoder to be deterministic, i.e., $Q_{Z|X=x} = \delta_{q(x)}$ for a function $q : \mathcal{X} \to \mathcal{Z}$. Then problem (P) reduces to

$$\min_{q:\mathcal{X}\to\mathcal{Z}} \mathbb{E}_{X\sim P_X} |X - g(q(X))|^2 + \lambda D(q_\sharp P_X, P_Z). \tag{S15}$$

The deterministic encoder q in (S15) was parameterized using a two-layer perceptron (MLP) with 32 hidden nodes and sigmoid activation functions. Sigmoid activation was also applied to the output layer in order to match the support of P_Z (e.g., for the KL divergence). The critic of the f-GAN or WGAN was also parameterized by a three-layer MLP with 32 hidden nodes and sigmoid activation functions. Batches of size 256 were sampled from P_X every iteration. One epoch was set to 150 iterations of updating the encoder q. The optimization of the auxiliary variable t in (7) was performed via SGD concurrently with the encoder. The encoder q was updated once after updating the critic ten times and the auxiliary variable five times. All optimization problems were solved using the RAdam optimizer (Liu et al., 2019) with momentum parameters $\beta = (0.5, 0.999)$. The initial learning rates were set to 0.0005 for the encoder and the critic, and 0.001 for the auxiliary variable t. For the training schedule, the learning rates were decayed by a factor of 0.8 every 50 epochs. We fix g to isolate the inner marginal-penalized primal OT problem and compare its optimized reconstruction value with known OT ground truth.

Threshold for the penalty coefficient. Let $c = d_{\mathcal{X}}^2$ and γ denote the Lipschitz constant of the decoder g. Recall that, for $D = \sqrt{D_f}$, the theoretical threshold in Corollary 3.8 with $p = 2$ is given by

$$\lambda^*(g) = \frac{L_p(g)}{C} = 2\text{diam}(\mathcal{X}) \max\{1, \gamma\} \text{diam}(\mathcal{Z}) / \sqrt{2f''(1)}.$$

In the case of the $D = W_1$, the denominator $\sqrt{2f''(1)}$ in the threshold formula is omitted. In the one-dimensional setting, since the decoder is simply the inverse cumulative distribution function of a truncated normal, we can calculate its Lipschitz constant and even calculate the threshold. We used this threshold as the penalty coefficient in the experiment. When we employ the annealing, the λ started from this threshold and was increased by 0.5 for each epoch.

Statistical test for the distributional constraint. In practice, the penalty $D(Q_Z, P_Z)$ hardly vanishes due to the stochastic nature of the model fitting procedure. To monitor whether the constraint $Q_Z = P_Z$ was satisfied during training, we employed an *equivalence test* (Wellek, 2010). The null and alternative hypotheses were formulated as

$$H_0 : \sup_z |F_Q(z) - F_P(z)| \geq \Delta \quad \text{vs.} \quad H_A :< \Delta,$$

where F_Q and F_P are the CDFs of Q_Z and P_Z, respectively. We reject the null hypothesis H_0 (conclude equivalence) if the upper confidence bound U of the KS statistic is less than a tolerance Δ. The upper confidence bound was constructed using the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Massart, 1990), which states that for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_z |F_n(z) - F(z)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

where F_n is the empirical CDF of n samples from CDF F. Let $F_{Q,n}$ be the empirical CDF from n samples of Q_Z. We compute the observed statistic, $D_{\text{obs}} = \sup_z |F_{Q,n}(z) - F_P(z)|$, and define the upper bound $U := D_{\text{obs}} + R_n(\alpha)$, where $R_n(\alpha) = \sqrt{\log(2/\alpha)/(2n)}$. The DKW inequality ensures that $\mathbb{P}\left(\sup_z |F_Q(z) - F_P(z)| \leq U\right) \geq 1 - \alpha$, which guarantees Type I error control

$$\sup_{Q' : \sup_z |F_{Q'}(z) - F_P| \geq \Delta} \mathbb{P}_{Q'}\left(U < \Delta\right) \leq \mathbb{P}\left(U < \sup_z |F_{Q'}(z) - F_P(z)|\right) \leq \alpha.$$

Since the multiple tests are performed during the training, we applied Bonferroni correction. For K planned tests, the significance level was adjusted to $\alpha' = \alpha/K$. The tolerance parameter Δ was set to be strictly greater than $R_n(\alpha')$ to ensure the test has non-zero power. The equivalent test was performed every 10 epochs. To be conservative, we terminate the training only if H_0 is rejected in at least three out of the last five tests.

Simulation results. We compare penalties on the exact side of the population theory, including square-root f-divergences and W_1, with f-divergences and squared MMD on the inexact side. Under finite parameterizations and numerical optimization, the unrooted penalties tend to underestimate the OT ground truth, whereas the square-root penalties reach its numerical neighborhood more closely, as illustrated in Figure S5.

Consistent with these observations, Table S4 shows that, under fixed coefficients, the square-root penalties generally yield reconstruction estimates closer to the ground truth and meet the equivalence tolerance earlier than their unrooted counterparts. Annealing reduces the reconstruction bias of several D_f penalties, qualitatively consistent with asymptotic recovery (Theorem 3.3) as the coefficient increases.

Sweep over the penalty coefficient. We swept λ over $2^0, 2^1, \ldots, 2^{12}$ with five seeds at each coefficient, for JS, KL, and MMD^2 on the inexact side and $\sqrt{JS}$ and $\sqrt{KL}$ on the exact side. Figure 2 reports the gap and Figure S4 the trained reconstruction itself.

If μ^λ minimizes $\mathcal{R} + \lambda\mathcal{D}$ over A, then $\lambda \mapsto \mathcal{R}(\mu^\lambda)$ is nondecreasing (cf. Dolgopolik (2016, Proposition 1)), and for an exact penalty, $\mathcal{R}(\mu^\lambda) = W_c$ if $\lambda > \lambda^*$. Empirically, the square-root curves enter the numerical neighborhood of the ground truth around $\lambda = 4$, below the available sufficient bound. This is compatible with the bound being sufficient rather than necessary, but does not identify the population exactness threshold. In addition, the exact penalties pass the equivalence test even at small coefficients, whereas the inexact penalties satisfy the test only when λ becomes sufficiently large. We note that passing the equivalence test indicates matching within the prescribed tolerance rather than exact equality.

Total variation distance. We do not include total variation in Table S4, although it is exact for the population marginal-penalized problem under the bounded-domain assumptions, because it is difficult to estimate reliably. Indeed, the total variation distance is also an f-divergence with $f(x) = |x - 1|$. However, with this nonsmooth f, the supremum in the variational representation of the f-divergence in Appendix A cannot be attained (Nowozin et al., 2016). This implies that estimating the total variation distance by f-GAN is very difficult, if not impossible.

Table S4: Full simulation results with various penalties. Epochs means the number of epochs until the model passed the equivalence test (at most 1000 in 1D and 500 in 2D). Rel. err. is the relative error (%) between reconstruction error and $W_2^2(P_X, P_G)$. Within each matched pair $(D_f, \sqrt{D_f})$, boldface marks the smaller absolute relative error. Underlining marks the smallest absolute relative error in each column. The W_1 and MMD^2 penalties have no matched counterpart and are eligible only for the second comparison.

Dim.	Penalty	Epochs		Rel. err. (%)	
		Fixed	Anneal	Fixed	Anneal
1D	KL	1000	842 (± 248)	-1.78 (± .100)	-0.340 (± .152)
	$\sqrt{\text{KL}}$	812 (± 266)	682 (± 310)	**-0.158** (± .223)	**-0.192** (± .231)
	rev-KL	1000	670 (± 147)	-1.79 (± .127)	**-0.161** (± .148)
	$\sqrt{\text{rev-KL}}$	526 (± 263)	370 (± 100)	**-0.138** (± .200)	-0.351 (± .268)
	χ^2	1000	678 (± 139)	-1.23 (± .143)	**-0.256** (± .185)
	$\sqrt{\chi^2}$	576 (± 185)	520 (± 190)	**-0.161** (± .084)	-0.293 (± .227)
	rev-χ^2	1000	842 (± 119)	-1.26 (± .169)	-0.347 (± .275)
	$\sqrt{\text{rev-}\chi^2}$	550 (± 204)	910 (± 111)	**-0.142** (± .207)	**-0.139** (± .107)
	Hellinger^2	1000	644 (± 195)	-2.25 (± .159)	-0.476 (± .272)
	Hellinger	822 (± 161)	762 (± 182)	**-0.135** (± .272)	**-0.009** (± .284)
	JS	1000	988 (± 24)	-2.32 (± .107)	-0.527 (± .325)
	$\sqrt{\text{JS}}$	710 (± 311)	830 (± 274)	**-0.251** (± .425)	**-0.104** (± .420)
	W_1	1000	1000	+0.148 (± .252)	+0.015 (± .196)
	MMD^2	1000	1000	-1.10 (± .039)	+0.420 (± .054)
2D	Hellinger^2	428 (± 90)	358 (± 129)	-2.06 (± .072)	**-0.339** (± .173)
	Hellinger	336 (± 164)	270 (± 189)	**-0.077** (± .215)	-0.421 (± .445)
	JS	428 (± 90)	358 (± 129)	-2.29 (± .099)	-0.319 (± .142)
	$\sqrt{\text{JS}}$	336 (± 164)	400 (± 97)	**-0.093** (± .160)	**-0.168** (± .223)
	W_1	340 (± 151)	396 (± 148)	-0.073 (± .258)	-0.591 (± 1.04)
	MMD^2	500	400 (± 123)	-6.60 (± .060)	-0.717 (± .158)

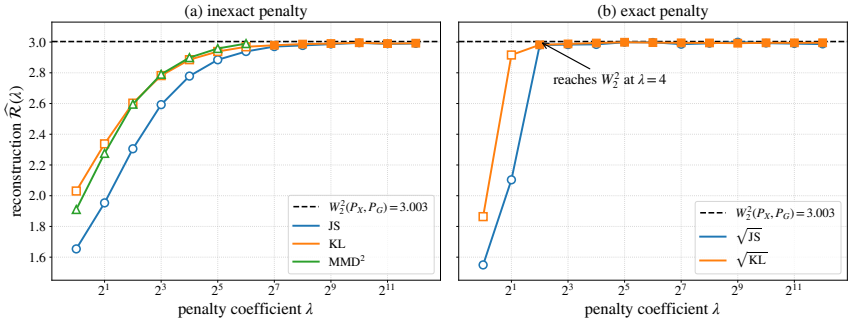


Figure S4: Trained reconstruction values in the 1D coefficient sweep of Figure 2. The horizontal reference marks the ground-truth value $W_2^2(P_X, P_G) = 3.003$.

D.2 Two-dimensional simulation

Setup. The support and parameters of P_X were randomly initialized. The P_X was defined as a bivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with mean $\mu = [-0.5545, -0.4041]^\top$ and covariance $\Sigma = \begin{pmatrix} 1.063 & -0.002955 \\ -0.002955 & 1.676 \end{pmatrix}$ truncated to the rectangular support $[-2.536, 1.416] \times [-2.424, 1.634]$.

To construct a two-dimensional problem with computable OT ground truth, we invoke Brenier's theorem (Villani, 2009). Since P_X is absolutely continuous with respect to Lebesgue measure, the quadratic-cost optimal transport to P_G is induced by a unique map $\nabla\varphi$, where φ is convex. Thus, given a convex function φ, the model distribution is constructed as a pushforward measure $P_G = (\nabla\varphi)_\sharp P_X$. To express this constructed primal OT instance through the WAE identity, we must write its model marginal as $g_\sharp P_Z$. Assuming the bijectivity of the decoder g provides a straightforward bridge to achieve this. This bijectivity guarantees that $(\nabla\varphi)_\sharp P_X = (g \circ g^{-1} \circ \nabla\varphi)_\sharp P_X = g_\sharp((g^{-1} \circ$

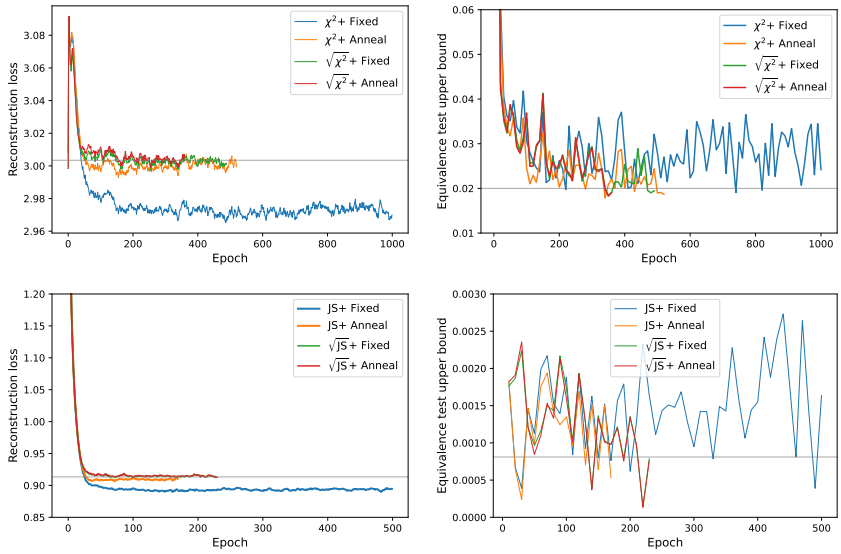


Figure S5: Training results for the f-divergences and their square-root penalties. The top row shows the 1D simulation using the χ^2 divergence, while the bottom row presents the 2D simulation using the JS divergence. Horizontal lines in both plots represent the true Wasserstein distance (left) and the rejection threshold for the Equivalence test (right). The reconstruction loss is smoothed for clarity, while raw values are presented for the equivalence test upper bound to emphasize threshold crossings.

$\nabla\varphi)_\sharp P_X)$. By establishing this relation, we can define our prior distribution P_Z as $(g^{-1} \circ \nabla\varphi)_\sharp P_X$. Consequently, we obtain

$$W_2^2(P_X, (\nabla\varphi)_\sharp P_X) = W_2^2(P_X, g_\sharp((g^{-1} \circ \nabla\varphi)_\sharp P_X)) = W_2^2(P_X, g_\sharp P_Z).$$

This equivalence, combined with the uniqueness of the OT map, rigorously implies that the optimal encoder is deterministic and given by $g^{-1} \circ \nabla\varphi$. Therefore, we modeled the encoder as a deterministic map $q : \mathcal{X} \to \mathcal{Z}$. The ground truth $W_2^2(P_X, (\nabla\varphi)_\sharp P_X) = 0.9134$ was approximated by $\frac{1}{n}\sum_{i=1}^n \|X_i - \nabla\varphi(X_i)\|_2^2$ where $X_i \overset{\text{i.i.d.}}{\sim} P_X$.

The Brenier potential φ and the bijective decoder g were parameterized by a fully connected input convex neural network (ICNN, Amos et al., 2017) and a RealNVP (Dinh et al., 2016) with coupling layers, respectively. Both were randomly initialized and fixed during training. The ICNN and the decoder network consisted of two convex layers and three coupling layers, respectively. To ensure strong convexity of the Brenier potential, a quadratic term, $\|x\|_2^2$, was added to the ICNN output. The encoder was parameterized by the pre-initialized φ followed by a RealNVP with the same architecture as the decoder; only the latter was trained. Since the network architecture does not guarantee overlapping supports for $q_\sharp P_X$ and P_Z, we restricted f-divergences to JS and squared Hellinger.

For optimization, we employed RAdam with learning rate 10^{-5}. The encoder was updated once after every ten critic updates and five updates of the auxiliary variable t. We initialized λ from a plug-in scale obtained by replacing the latent diameter and decoder Lipschitz constant in Corollary 3.8 with the maximum sampled pairwise distance and maximum sampled Jacobian norm. These sampled extrema need not bound their global counterparts, so this initialization is a heuristic rather than a sufficient threshold or certificate. In the annealing experiment, λ was subsequently increased by 1.0 at each epoch.

Testing distributional equality. In the 2D simulation, the hypotheses for the equivalence test are formulated as follows

$$H_0 : \text{MMD}^2(Q_Z, P_Z) \geq \Delta \quad \text{vs.} \quad H_A : \text{MMD}^2(Q_Z, P_Z) < \Delta.$$

Owing to the computational complexity of the KS statistic in higher dimensions, we adopted the block MMD test (Zaremba et al., 2013). The batch estimator of squared MMD, denoted by $\hat{\eta}$, was

Table S5: Ablation study over fixed penalty coefficients λ on CelebA with the square root of the JS divergence under a logit normal prior.

λ	FID	Density	Coverage	Precision	Recall
1	78.0	21.7	1.000	0.957	0.003
10	58.2	4.49	0.999	0.921	0.185
100	56.9	1.06	0.969	0.878	0.690
1000	58.5	0.902	0.950	0.836	0.694
10000	57.4	4.08	0.999	0.923	0.193

computed by averaging the unbiased estimator of squared MMD across blocks. Invoking the Central Limit Theorem (CLT), $\hat{\eta}$ asymptotically follows a normal distribution with mean $\mathrm{MMD}^2(Q_Z, P_Z)$ provided that the latter is strictly positive (Zaremba et al., 2013, Theorem 1). Following the Z-test framework, we define the upper confidence bound as $U := \hat{\eta} + z_{1-\alpha}\hat{\sigma}_{\hat{\eta}}$ where $z_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal and $\hat{\sigma}_{\hat{\eta}}$ is the standard error of the block statistics. Then the null hypothesis H_0 is rejected if $U < \Delta$. The tolerance Δ was determined by taking the mean of $\hat{\eta}$ plus twice the standard error, estimated from two independent samples drawn from P_Z. We utilized a mixture of Gaussian kernels, with bandwidths determined by the median heuristic and its multiples.

Simulation results. The penalties that are exact at the population level under our assumptions, namely $\sqrt{\mathrm{JS}}$, Hellinger, and W_1, generally meet the equivalence tolerance earlier and produce reconstruction estimates closer to the ground truth in this experiment. The standard f-divergences and MMD^2 yield lower reconstruction estimates under the fixed coefficients, even in runs that pass the statistical equivalence test.

D.3 Latent Space Coverage Analysis on Image Dataset

The encoder and decoder architectures comprise four ResNet layers integrated with fully connected layers to map to and from a 64-dimensional latent space. The model was trained for 1000 epochs with a batch size of 512 using mixed-precision training (bf16-mixed) to optimize computational efficiency and memory usage. During training, we performed three updates for the discriminator for every single update of the autoencoder. We employed the AdamW optimizer (Loshchilov & Hutter, 2019) for the autoencoder with an initial learning rate 8×10^{-4}, while the discriminator was optimized using Adam (Kingma & Ba, 2015) with a 2.5 times higher learning rate to facilitate stable adversarial training. In the case of the square-root penalty formulation, an auxiliary variable was optimized concurrently using the Adam optimizer with a learning rate of 8×10^{-3}. We fixed $\lambda = 100$ in all CelebA experiments. This value was selected in a preliminary sweep over $\lambda \in \{1, 10, 100, 10^3, 10^4\}$ using $\sqrt{\mathrm{JS}}$ under the logit-normal prior. Among the tested values, $\lambda = 100$ was the smallest for which Density and Coverage approached their $Q_Z = P_Z$ reference values while FID remained comparable (Table S5). This is an empirical selection based on finite-sample diagnostics, rather than a verification of $Q_Z = P_Z$ or of population exactness.

Density, Coverage, Precision, and Recall metric. The Density/Coverage metrics (Naeem et al., 2020) and the Precision/Recall metrics (Kynkäänniemi et al., 2019) measure the alignment and overlap between the manifolds of real and synthetic samples. Let $\{Z_i\}_{i=1}^N$ be N samples from the

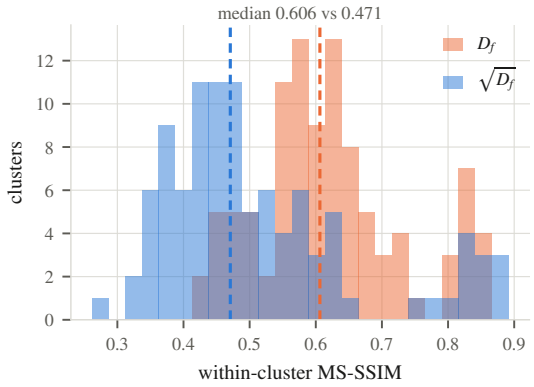


Figure S6: Local image similarity under the logit-normal prior, comparing JS and $\sqrt{JS}$. Within-cluster MS-SSIM in the densest 1% of aggregate-posterior neighborhoods.

prior P_Z, and $\{Z'_j\}_{j=1}^N$ be samples from the aggregate posterior Q_Z. These metrics are defined as

$$\text{density} := \frac{1}{kN} \sum_{j=1}^{N} \sum_{i=1}^{N} I(Z'_j \in \text{NBD}_k(Z_i)),$$

$$\text{coverage} := \frac{1}{N} \sum_{i=1}^{N} I(\exists j \text{ s.t. } Z'_j \in \text{NBD}_k(Z_i)),$$

$$\text{precision} := \frac{1}{N} \sum_{j=1}^{N} I(\exists i \text{ s.t. } Z'_j \in \text{NBD}_k(Z_i)),$$

$$\text{recall} := \frac{1}{N} \sum_{i=1}^{N} I(\exists j \text{ s.t. } Z_i \in \text{NBD}_k(Z'_j)),$$

where $I(\cdot)$ is the indicator function. The neighborhood $\text{NBD}_k(x)$ is defined as the ball centered at x with radius equal to the distance to its k-th nearest neighbor within its respective sample set (i.e., $\{Z_i\}_{i=1}^N$ for real samples, and $\{Z'_i\}_{i=1}^N$ for synthetic samples). Following the recommendations of Naeem et al. (2020), we used $k = 5$ and $N = 20260$ in our experiments. These values yield an expected Coverage value of approximately 0.969, which exceeds the standard baseline of 0.95, and an expected Density value of 1.0.

Local MS-SSIM diagnostic. We use intra-neighborhood Multi-Scale Structural Similarity Index Measure (MS-SSIM, Wang et al., 2003) as a diagnostic of local image redundancy. Values range from 0 to 1, with higher values indicating greater similarity.

We apply the procedure separately to the D_f and $\sqrt{D_f}$ models: for each model, we select the densest 1% of aggregate-posterior neighborhoods by k-NN radius, decode the $k + 1$ points in each neighborhood, and average their pairwise MS-SSIM. Because the two models generally select different neighborhoods, the resulting distributions are descriptive rather than a matched or causal comparison. Lower MS-SSIM indicates less visual redundancy within the selected neighborhoods, but by itself neither rules out mode collapse nor establishes semantic diversity. Figure S6 shows that the $\sqrt{D_f}$ diagnostic is shifted toward lower values in this experiment.

Discussion on the standard normal prior. We include the standard normal prior because the population exactness results for $\sqrt{\text{KL}}$ and $\sqrt{\chi^2}$ allow unbounded latent support under the assumptions of Section 3.4; the experiment does not certify exactness for the trained networks. Unlike the logit-normal results, the standard-normal results show no uniform preference. Standard D_f gives Density closer to 1, whereas $\sqrt{D_f}$ gives Coverage closer to 0.969; FID, Precision, and Recall are comparable within each pair. The empirical effect of the penalty is therefore prior-dependent; these metrics alone do not identify the geometric cause of the difference.

D.4 The Penalty Coefficient on CelebA

For $p = 2$ and $D = \sqrt{\mathrm{JS}}$, Corollary 3.8 gives the sufficient threshold $2\,\mathrm{diam}(\mathcal{X})\max\{1,\gamma\}\,\mathrm{diam}(\mathcal{Z})$. For the logit-normal experiment on CelebA, $\mathcal{X} = [-1,1]^{3\times 64\times 64}$ and $\mathcal{Z} = [0,1]^{64}$, hence $\mathrm{diam}(\mathcal{X}) = 128\sqrt{3}$ and $\mathrm{diam}(\mathcal{Z}) = 8$. To illustrate the numerical scale of this expression, we replace the Lipschitz constant of g by the largest Jacobian norm observed over prior samples, $\widehat{\gamma} \approx 550$, which gives the plug-in value 1.95×10^6. This quantity is neither a certified sufficient threshold nor an estimate of the minimal exact coefficient, since $\widehat{\gamma}$ may not upper-bound the global Lipschitz constant.

Even this non-certified plug-in scale exceeds the empirical choice $\lambda = 100$ by four orders of magnitude, illustrating the conservatism of the ambient diameter and worst-case Lipschitz factors. The bound in Corollary 3.8 uses global worst-case quantities to establish the existence of a finite population-level exact coefficient; it is not intended as a calibration rule for neural training. We therefore select $\lambda = 100$ by the preliminary ablation described in Table S5. Indeed, classical exact penalty theory explicitly cautions against choosing an unnecessarily large penalty parameter (e.g., Nocedal & Wright, 2006, Chapter 17). An excessively large coefficient severely impairs the conditioning of the penalized objective, complicating numerical optimization and amplifying finite-precision and stochastic gradient errors.

E Summary of Assumptions

Table S6 provides a comprehensive summary of the formal assumptions and theoretical guarantees established for each main result. All guarantees concern the population P_Z-marginal penalized primal OT problem and its global minimizers over all joint laws with fixed P_X-marginal, equivalently the unrestricted encoder class $\mathcal{Q}^*$.

F Detailed Comparison with Prior Work

Table S7 compares our contributions with exact-penalty theory and prior analyses of the WAE-induced marginal-penalized primal OT formulation.

Table S6: Scope, assumptions, and theoretical guarantees of the main results.

Result	Penalty	Main Assumptions	Guarantee
Theorem 3.3	Weakly l.s.c. D (e.g., D_f, W_p, MMD_k^2)	$\mathcal{Z} \times \mathcal{X}$ compact, c and g continuous, $\lambda_k \uparrow \infty$	Subsequential convergence of penalized minimizers to a constrained minimizer
Proposition 3.4	D_f with $f'(1)$ existing, or MMD_k^2 with bounded continuous k	$\mathcal{X}$, $\mathcal{Z}$ bounded, $\mathrm{supp}(P_Z) = \mathcal{Z}$, $c(\cdot, g(\cdot))$ continuous and bounded, active constraint $\Delta < 0$	Strict inexactness: $V_\lambda < W_c$ for all $\lambda > 0$
Theorem 3.5	D_f; characteristic MMD_k^2	Assumptions of Proposition 3.4; $f \in C^2$ and $f''(1) > 0$ for the D_f lower bound; generalized Pinsker additionally for the matching upper bound	For D_f, the C^2 assumptions give $W_c - V_\lambda \geq \min\{c_1/\lambda, \lvert\Delta\rvert/8\}$, while generalized Pinsker gives $W_c - V_\lambda \leq c_2/\lambda$ and hence $\Theta(1/\lambda)$; for MMD^2, $W_c - V_\lambda \geq \min\{c_{\mathrm{M}}/\lambda, \lvert\Delta\rvert/4\}$
Corollary 3.8	$\sqrt{D_f}$ (generalized Pinsker), TV, W_1	$c = d_{\mathcal{X}}^p$ with $p \geq 1$, $\mathcal{X}$ and $\mathcal{Z}$ bounded, g γ-Lipschitz	Exactness: minimizers of (P′) and (O′) coincide for $\lambda > \lambda^*(g) = L/C$
Lemma 3.2, $p = 1$	As in Corollary 3.8	$c = d_{\mathcal{X}}$, finite first moments, $\mathcal{Z}$ bounded, g Lipschitz	Boundedness of $\mathcal{X}$ is relaxed
§3.4	$\sqrt{\mathrm{KL}}$	P_Z satisfies $T_1(\tau)$ (e.g., Gaussian), $\mathcal{X}$ bounded unless $p = 1$, g Lipschitz	Exactness with $C = 1/\sqrt{2\tau}$; $\mathcal{Z}$ unbounded
Proposition 3.9	$\sqrt{\chi^2}$	$\mathbb{E}_{P_Z}[\lVert Z\rVert^2] < \infty$, $\mathcal{X}$ bounded unless $p = 1$, g Lipschitz	Exactness with $C = (\mathbb{E}_{P_Z}[\lVert Z\rVert^2])^{-1/2}$; $\mathcal{Z}$ unbounded
Proposition 4.2	Any D_e satisfying condition (3)	$c = d_{\mathcal{X}}^p$, $\mathcal{X}$ bounded with diameter K, decoders uniformly γ-Lipschitz, $\lambda > pK^{p-1}\max\{1, \gamma\}/C$	Uniform empirical deviation bounded by $pK^{p-1}W_p(P_X, \hat{P}_X^n)$ without capacity terms for $\mathcal{G}$
Proposition 4.3	Any exact penalty D_e	Assumptions of Proposition 4.2, empirical minimization over $\mathcal{G}$; for rate, $P_X = g_\sharp^\star P_Z$ with P_Z on compact d-manifold, $d \geq 2p$	Distributional W_p oracle inequality and rate $\mathcal{O}(n^{-1/s})$ for all $s > d$

Table S7: Detailed comparison between prior literature and our contributions.

Prior Work	Established Results	Open Questions / Limitations	Our Contributions
Dolgopolik (2016)	Abstract sufficient conditions for exact penalties in metric spaces	the geometry of the one-sided marginal constraint in the WAE-induced primal OT formulation, which marginal penalties satisfy the conditions, and whether finite exactness can fail	Metric space formulation on $\mathbb{W}_1(\mathcal{Z} \times \mathcal{X})$ via (4), the objective bound of Lemma 3.2, explicit divergence constants (Theorem 3.7), and inexactness with a rate (Proposition 3.4 and Theorem 3.5)
Husain et al. (2019)	Equivalence between (O) and (P) for D_f under metric cost $c = d_{\mathcal{X}}$ when $\lambda > \sup_{P'} W_1/D_f$, and generalization bounds	Finiteness of this threshold for D_f, uniform generalization guarantees under general p-th power costs $c = d_{\mathcal{X}}^p$ ($p > 1$)	Shows that the ratio-based threshold is infinite for differentiable f at 1 and, more strongly, proves that no finite coefficient recovers the constrained problem under an active marginal constraint (Proposition 3.4); establishes finite thresholds for $\sqrt{D_f}$, derives uniform generalization bounds under general costs $c = d_{\mathcal{X}}^p$ with no complexity term of $\mathcal{G}$ (Proposition 4.2)
Patrini et al. (2020)	Exactness of W_p penalty for the rooted objective (1) under non-atomic P_X	Standard p-th power objectives, and necessity of non-atomicity	Recovers (1) without non-atomicity assumptions (Appendix C); establishes exactness for $\sqrt{KL}$, $\sqrt{\chi^2}$, and W_1 under unbounded priors
Chakrabarty & Das (2021); Chakrabarty et al. (2025)	Error bounds of order $\mathcal{O}(n^{-1/s} + t)$ for all $s > d$ and a relaxed objective with latent tolerance t	Eliminating the non-vanishing tolerance parameter t	Under population exactness, eliminates the tolerance parameter and achieves rate $\mathcal{O}(n^{-1/s})$ for all $s > d$ (Proposition 4.3)
Chakraborty & Bartlett (2024)	Excess risk bounds of order $\mathcal{O}(n^{-1/(d+2)})$ restricted to the penalized surrogate loss	Establishing direct convergence rates for the generated distribution under W_1	Establishes a distribution-level W_1 oracle inequality and achieves convergence rate $\mathcal{O}(n^{-1/s})$ for all $s > d$ (Proposition 4.3)