

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 GEOMETRIC IB: IMPROVING INFORMATION BOTTLE- NECK WITH GEOMETRY-AWARE COMPRESSION ON STATISTICAL MANIFOLDS

Anonymous authors

Paper under double-blind review

ABSTRACT

We revisit the Information Bottleneck (IB) through the lens of information geometry and propose a Geometric Information Bottleneck (G-IB) that dispenses with direct mutual information (MI) estimation. We show that mutual information $I(X; Z)$ and $I(Z; Y)$ admit exact projection forms as minimal Kullback–Leibler (KL) distances from the joint distributions to their respective independence manifolds. Guided by this view, G-IB controls information compression with two complementary terms: (i) a distribution-level Fisher–Rao (FR) discrepancy, which matches KL to second order and is reparameterization-invariant; and (ii) a geometry-level Jacobian–Frobenius (JF) term that provides a local capacity-type upper bound on $I_\phi(Z; X)$ by penalizing pullback volume expansion of the encoder. We further derive a natural-gradient optimizer consistent with the FR metric and prove that the standard additive natural-gradient step is first-order equivalent to the geodesic update. We conducted extensive experiments and observed that the G-IB achieves a better trade-off between prediction accuracy and compression ratio in the information plane than the mainstream IB baselines on popular datasets. G-IB offers a principled and scalable alternative that unifies distributional and geometric regularization under a single bottleneck multiplier, improving invariance and optimization stability. The source code of G-IB is released at <https://anonymous.4open.science/r/G-IB-0569>.

1 INTRODUCTION

The Information Bottleneck (IB) principle (Tishby et al., 2000) casts representation learning as extracting a representation Z from X that preserves only what is useful for predicting Y . Concretely, one seeks an encoder $q_\phi(z | x)$ such that Z carries as much information about Y as possible while remaining maximally compressed with respect to X , which can be formulated as:

$$\max_{\phi} I_\phi(Z; Y) \quad \text{s.t.} \quad I_\phi(X; Z) \leq R, \quad (1)$$

where $I(\cdot; \cdot)$ denotes the mutual information, ϕ are the parameters of the encoder, and R sets the compression budget. Here, $I_\phi(\cdot; \cdot)$ is computed under the data distribution $p(x, y)$ and the encoder $q_\phi(z | x)$. To address this constrained optimization problem, the IB method ((Tishby et al., 2000; Alemi et al., 2016)) introduces a positive Lagrange multiplier β , transforming the problem into

$$\min_{\phi} -I_\phi(Z; Y) + \beta I_\phi(X; Z), \quad (2)$$

where $\beta \geq 0$ (the “bottleneck” parameter) balances predictive sufficiency against compression.

The IB principle is appealing because it formalizes what constitutes a useful representation via a fundamental balance between compression and predictive sufficiency (Alemi et al., 2016; Tishby et al., 2000; Wu et al., 2020a). Thus, IB and a wide range of variants (Wan et al., 2021; Yang et al., 2025; Yu et al., 2024; Zhai & Zhang, 2022) have been adopted across diverse applications, including image segmentation (Xu et al., 2024), domain generalization (Li et al., 2022), semantic communication (Xie et al., 2023; Wang et al., 2024), and privacy compression (Dubois et al., 2021; Razeghi et al., 2023). Moreover, prior work (Shwartz-Ziv & Tishby, 2017) suggests that IB provides a principled lens for interpreting certain training dynamics of deep neural networks and unveil universal attrition to interpret vision transformers (Hong et al., 2025).

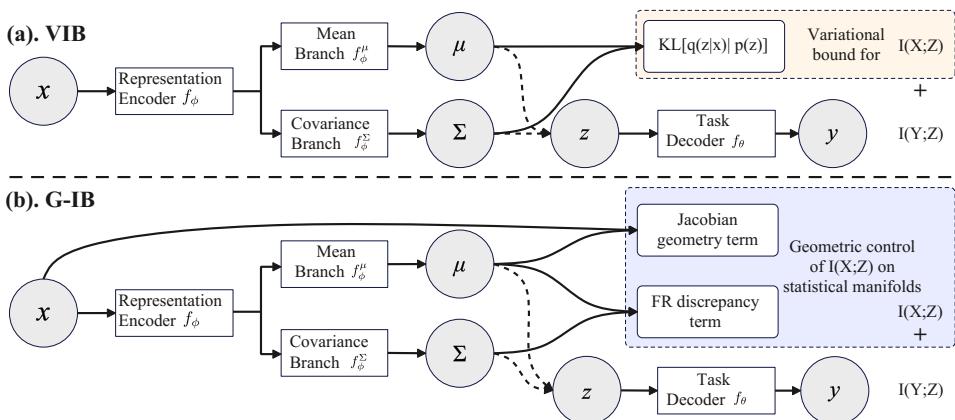


Figure 1: Comparison of VIB and G-IB. Both models parametrize the encoder as $q_\phi(z \mid x) = \mathcal{N}(\mu(x), \text{diag}(\sigma^2(x)))$ by a network f_ϕ and use a task decoder $p_\theta(y \mid z)$ to increase $I(Z; Y)$ by a network f_θ . (a) VIB: compression is enforced by the variational upper bound. (b) G-IB: replaces explicit MI estimation with two geometry-aware penalties computed deterministically on statistical manifolds: a Fisher–Rao quadratic proxy \mathcal{L}_{FR} and a Jacobian–Frobenius term \mathcal{L}_{JF} . Solid arrows denote deterministic mappings; dashed arrows indicate reparameterized sampling $z = \mu + \sigma \odot \varepsilon$.

Research Gap. Despite this success, most practical IB implementations optimize Euclidean surrogates of mutual information (MI), e.g., variational bounds as in VIB (Alemi et al., 2016) or neural MI estimators such as MINE (Belghazi et al., 2018). These surrogates disregard the statistical-manifold geometry of the encoder or posterior family and thus offer no explicit geometric guarantees (e.g., reparameterization invariance, curvature-aware regularization) over $I(X; Z)$. Recent extensions, such as structure IB (Yang et al., 2025; Hu et al., 2024), try to extract the structure information from the input. However, they still rely on Euclidean MI proxies, often degrading accuracy in the strong-compression regime and making results highly sensitive to β and estimator hyperparameters.

Research Question. This motivates the following research question: “Can we design a geometry-aware IB that operates on statistical manifolds and provides stable, principled control of representation compression?”

In this paper we introduce the Geometric Information Bottleneck (G-IB), which reframes IB through the lens of statistical–manifold geometry. We first establish exact projection characterizations: both $I(X; Z)$ and $I(Z; Y)$ can be written as minimal Kullback–Leibler (KL) distances from the corresponding joint distributions to their independence submanifolds. Then, we design the G-IB method, which regulates compression via two complementary components: (i) a *distribution-level* Fisher–Rao (FR) discrepancy that agrees with KL to second order and is invariant under smooth reparameterizations of z ; and (ii) a *geometry-level* Jacobian–Frobenius (JF) penalty that yields a local capacity–type upper bound on $I_\phi(Z; X)$ by discouraging pullback volume expansion of the encoder. Finally, we derive an optimizer with respect to the Fisher–Rao (FR) metric whose update direction is the natural gradient, and we prove first-order equivalence with the geodesic update. To summarize, we make the following contributions:

- **Geometric Reformulation of IB.** We show that both $I(Z; X)$ and $I(Z; Y)$ admit exact projection forms as minimal KL distances from the joint distributions to their respective independence manifolds, clarifying the geometric structure underlying the IB principle.
- **A G-IB Solution.** We propose G-IB, which controls compression via two complementary terms: (i) a distribution-level Fisher–Rao (FR) discrepancy and (ii) a geometry-level Jacobian–Frobenius (JF) penalty. We also derive a natural-gradient optimizer consistent with the FR metric and prove that: the standard additive natural-gradient step is first-order equivalent to the geodesic (exponential-map) update.
- **Empirical Validation.** We conducted extensive experiments to compare with representative benchmarks. G-IB attains favorable accuracy–compression trade-offs in the information plane relative to the state-of-the-art IB baselines, with improved robustness in strong compression regimes.

108 As the page limitation, we provide the **Related Work** Section in Appendix B.
 109

110 2 PROBLEM STATEMENT FROM A GEOMETRIC VIEW

111 Let \mathcal{P} be the statistical manifold of all joint distributions over (X, Z) and let the independence
 112 manifold be $\mathcal{I}_{XZ} = \{q(x)r(z)\}$. In exponential (e-) coordinates \mathcal{I}_{XZ} is a e-flat submanifold
 113 (Amari & Nagaoka, 2000; Amari, 2016). For any q and r , we have the information-geometric
 114 Pythagorean identity (Amari & Nagaoka, 2000)

$$117 \text{KL}(p_\phi(x, z) \| q(x)r(z)) = \underbrace{\text{KL}(p_\phi(x, z) \| p(x)p_\phi(z))}_{=I_\phi(X; Z)} + \underbrace{\text{KL}(p(x)p_\phi(z) \| q(x)r(z))}_{\geq 0}, \quad (3)$$

118 whence $I_\phi(Z; X) = \min_{q,r} \text{KL}(p_\phi(x, z) \| q(x)r(z))$ with minimizer $(q, r) = (p(x), p_\phi(z))$. An
 119 identical relation holds for (Y, Z) by replacing x with y , q with q' , and r with r' :

$$122 \text{KL}(p_\phi(y, z) \| q'(y)r'(z)) = I_\phi(Z; Y) + \text{KL}(p(y)p_\phi(z) \| q'(y)r'(z)). \quad (4)$$

123 We assume absolute continuity so that all KL terms are finite and Fubini’s theorem (Kallenberg,
 124 1997) applies; in particular $p_\phi(x, z) \ll q(x)r(z)$ and $p_\phi(y, z) \ll q'(y)r'(z)$ for candidate product
 125 measures. We provide the detailed proof of Eq. (3) in Appendix C, and Eq. (4) can also be proved
 126 in the same way. The Information Bottleneck loss (Tishby et al., 2000), $\mathcal{L}_{\text{IB}}(\phi) = \beta I_\phi(Z; X) -$
 127 $I_\phi(Z; Y)$, can thus be written exactly as

$$128 \mathcal{L}_{\text{IB}}(\phi) = \beta \min_{q,r} \text{KL}(p_\phi(x, z) \| q(x)r(z)) - \min_{q',r'} \text{KL}(p_\phi(y, z) \| q'(y)r'(z)), \quad (5)$$

130 where the inner minima are achieved at $(q, r) = (p(x), p_\phi(z))$ and $(q', r') = (p(y), p_\phi(z))$. Opti-
 131 mizing Eq. (5) over ϕ is therefore equivalent to:

- 133 • *push* $p_\phi(x, z)$ *toward* the independence manifold \mathcal{I}_{XZ} by minimizing
 $\beta \text{KL}(p_\phi(x, z) \| p(x)p_\phi(z))$;
- 135 • *pull* $p_\phi(y, z)$ *away from* its independence manifold $\mathcal{I}_{YZ} = \{q'(y)r'(z)\}$ by maximizing
 $\text{KL}(p_\phi(y, z) \| p(y)p_\phi(z))$.

138 Although the projection-based formulation is exact, directly optimizing it is computationally chal-
 139 lenging due to the need to evaluate KL projections and encoder marginals at scale. We introduce
 140 an approximate solution that replaces explicit MI estimation with geometry-derived surrogates.

142 3 GEOMETRIC INFORMATION BOTTLENECK METHOD

144 **Overview.** As introduced in the above section, we recast the IB objective through information geo-
 145 metry: the mutual information $I_\phi(X; Z)$ and $I_\phi(Z; Y)$ are the minimal KL distances from the joint
 146 distributions $p_\phi(x, z)$ and $p_\phi(y, z)$ to their respective independence manifolds, turning the IB La-
 147 grangian into a difference of projection distances. Based on this, our Geometric IB (G-IB) controls
 148 compression with two complementary terms: (i) a *distribution-level* Fisher–Rao (FR) discrepancy
 149 and (ii) a *geometry-level* Jacobian–Frobenius (JF) term. During training, we propose a natural gra-
 150 dient descent method that combines the geometric information to achieve a better optimization effect.

152 3.1 DISTRIBUTION PROXY VIA THE FISHER–RAO (FR) QUADRATIC

153 We approximate the conditional–marginal divergence that defines the compression term by the local
 154 second-order FR metric:

$$156 I_\phi(Z; X) = \mathbb{E}_{p(x)} D_{\text{KL}}(q_\phi(z|x) \| p_\phi(z)) \approx \frac{1}{2} \mathbb{E}_{p(x)} d_{\text{FR}}(q_\phi(z|x), r(z))^2, \quad (6)$$

157 where $r(z)$ is a reference marginal (e.g., a standard normal or a learned prior), and $d_{\text{FR}}(\cdot, \cdot)$ denotes
 158 the Fisher–Rao geodesic distance between distributions, i.e., the Riemannian distance induced by
 159 the Fisher information metric. The approximation follows from the local equivalence for smooth
 160 parametric families $\{p_\theta\}$, for $\theta' = \theta + \Delta$ with $\|\Delta\|$ small:

$$161 D_{\text{KL}}(p_{\theta'} \| p_\theta) = \frac{1}{2} \Delta^\top F(\theta) \Delta + o(\|\Delta\|^2) = \frac{1}{2} d_{\text{FR}}(p_{\theta'}, p_\theta)^2 + o(\|\Delta\|^2), \quad (7)$$

162 with $F(\theta)$ the Fisher information. Replacing $p_{\theta'}$ by $q_\phi(z|x)$ and p_θ by $r(z)$ in Eq. (7) yields Eq. (6).
 163 We provide the proof in Appendix D.

164 **Diagonal Gaussian Example.** If $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \text{diag}(\sigma_\phi^2(x)))$ and $r(z) = \mathcal{N}(0, I)$, the exact
 165 KL is

$$167 \quad D_{\text{KL}}(q_\phi(z|x) \parallel \mathcal{N}(0, I)) = \frac{1}{2} \sum_{j=1}^{d_z} \left(\mu_j(x)^2 + \sigma_j(x)^2 - \log \sigma_j(x)^2 - 1 \right). \quad (8)$$

170 Expanding Eq. (8) at $\mu = 0, \sigma^2 = 1$ gives
 171

$$172 \quad D_{\text{KL}}(q_\phi \parallel \mathcal{N}(0, I)) = \frac{1}{2} \|\mu\|_2^2 + \frac{1}{4} \|\log \sigma^2\|_2^2 + O(\|\log \sigma^2\|_2^3), \quad (9)$$

173 which equals $\frac{1}{2} d_{\text{FR}}(q_\phi, \mathcal{N}(0, I))^2$ up to second order. In practice, when $r = \mathcal{N}(0, I)$ we can optimize
 174 the closed-form KL in Eq. (8): $\mathcal{L}_{\text{KL}}(\phi) = \widehat{\mathbb{E}}_x D_{\text{KL}}(q_\phi(z|x) \parallel r(z))$. If r is learned/complex
 175 (e.g., VampPrior/flow), we may optimize the FR proxy
 176

$$177 \quad \mathcal{L}_{\text{FR}}(\phi) = \frac{1}{2} \widehat{\mathbb{E}}_x d_{\text{FR}}(q_\phi(z|x), r(z))^2, \quad (10)$$

178 and optionally monitor $\widehat{I}_{XZ}^{\text{KL}} = \widehat{\mathbb{E}}_x D_{\text{KL}}(q_\phi(z|x) \parallel \widehat{p}_\phi(z))$ to gauge tightness.
 179

181 3.2 GEOMETRIC BOUND VIA THE JACOBIAN–FROBENIUS TERM

183 Assume a reparameterized encoder $z = f_\phi(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \Sigma(x))$, and denote the Jacobian
 184 $J_f(x) = \partial f_\phi(x) / \partial x$. The pullback metric on the input manifold is $g_x = J_f(x)^\top \Sigma(x)^{-1} J_f(x)$.
 185

186 **Local Capacity-type Upper Bound.** Linearizing f_ϕ around x and letting C_x be the local input
 187 covariance, a Gaussian channel upper bound yields

$$188 \quad I_\phi(Z; X) \leq \frac{1}{2} \mathbb{E}_{p(x)} \left[\log \det \left(I + \Sigma(x)^{-\frac{1}{2}} J_f(x) C_x J_f(x)^\top \Sigma(x)^{-\frac{1}{2}} \right) \right]. \quad (11)$$

191 Under a unit local energy constraint $C_x \preceq I$ (Loewner order) and the monotonicity of $\log \det$ on
 192 PSD cone \mathbb{S}_+^d , we obtain the pointwise bound

$$194 \quad I_\phi(Z; X) \leq \frac{1}{2} \mathbb{E}_{p(x)} \left[\log \det \left(I + \Sigma(x)^{-\frac{1}{2}} J_f(x) J_f(x)^\top \Sigma(x)^{-\frac{1}{2}} \right) \right] \quad (12)$$

$$195 \quad = \frac{1}{2} \mathbb{E}_{p(x)} \left[\log \det \left(I + J_f(x)^\top \Sigma(x)^{-1} J_f(x) \right) \right], \quad (13)$$

$$196 \quad \leq \frac{1}{2} \mathbb{E}_{p(x)} \text{Tr}(\Sigma(x)^{-1} J_f(x) J_f(x)^\top) = \frac{1}{2} \mathbb{E}_{p(x)} \|\Sigma(x)^{-\frac{1}{2}} J_f(x)\|_F^2 =: \frac{1}{2} \text{JF}(\phi). \quad (14)$$

200 *Proof sketch.* Since Z depends on X only through $(f_\phi(X), \Sigma(X))$, we have the Markov chain
 201 $X \rightarrow (f_\phi(X), \Sigma(X)) \rightarrow Z$ and thus $I(X; Z) = I((f_\phi(X), \Sigma(X)); Z)$. Conditioning on x and
 202 linearizing f_ϕ at x while holding $\Sigma(x)$ fixed locally yields a (local) linear Gaussian channel with
 203 gain $J_f(x)$ and noise covariance $\Sigma(x)$. Under a unit local energy constraint on the input covariance
 204 $C_x \preceq I$, the Gaussian maximizes entropy for fixed covariance, giving the log-det bound in Eq. (11).
 205 Using $\det(I + AB) = \det(I + BA)$ gives Eq. (13), and applying $\log \det(I + A) \leq \text{Tr}(A)$ for
 206 $A \succeq 0$ yields Eq. (14). \square

208 **Unbiased Hutchinson Estimator.** The trace in Eq. (14) can be estimated without forming explicit
 209 Jacobians. For any $v \sim \mathcal{N}(0, I_{d_x})$,

$$210 \quad \mathbb{E}_v [\|\Sigma(x)^{-1/2} J_f(x) v\|_2^2] = \text{Tr}(\Sigma(x)^{-1} J_f(x) J_f(x)^\top). \quad (15)$$

212 With S i.i.d. probe vectors $\{v_s\}_{s=1}^S$, define the per-sample estimator
 213

$$214 \quad \widehat{\text{JF}}(x) := \frac{1}{S} \sum_{s=1}^S \|\Sigma(x)^{-1/2} J_f(x) v_s\|_2^2, \quad \text{so that } \mathbb{E}_v [\widehat{\text{JF}}(x)] = \text{Tr}(\Sigma^{-1} J_f(x) J_f(x)^\top). \quad (16)$$

216 The training objective is the batch average
 217

$$\mathcal{L}_{\text{JF}}(\phi) := \widehat{\mathbb{E}}_{x \sim \text{batch}} [\widehat{\mathbf{JF}}(x)]. \quad (17)$$

219 In automatic differentiation frameworks, compute $J_f(x)v_s$ via Jacobian–vector products (JVP),
 220 costing $O(S)$ forward-mode calls per x ; $S = 1$ or 2 is typically sufficient.
 221

222 **Isotropic/diagonal noise.** If $\Sigma(x) = \sigma(x)^2 I$, then
 223

$$I_\phi(Z; X) \leq \frac{1}{2} \mathbb{E}_{p(x)} \log \det \left(I + \frac{1}{\sigma(x)^2} J_f J_f^\top \right) \leq \frac{1}{2} \widehat{\mathbb{E}}_x \frac{\|J_f(x)\|_F^2}{\sigma(x)^2}. \quad (18)$$

225 With a scalar floor $\sigma_{\min}^2 > 0$, this yields a simple, stable surrogate.
 226

227 *Geometric meaning.* Because $g_x = J_f^\top \Sigma^{-1} J_f$, we have $\text{Tr}(g_x) = \|\Sigma(x)^{-\frac{1}{2}} J_f(x)\|_F^2$, the direction-
 228 averaged local stretch (Dirichlet energy density under the Σ^{-1} metric). Minimizing the JF term
 229 therefore controls average geodesic-length distortion, providing a principled compression surrogate
 230 for $I_\phi(Z; X)$.
 231

232 3.3 NATURAL-GRADIENT OPTIMIZATION FOR G-IB

233 Building on the above, we formulate the G-IB objective as
 234

$$\mathcal{L}_{\text{G-IB}}(\phi, \theta) = \underbrace{\mathbb{E}_{p(x,y)} \mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(y | z)]}_{\uparrow I_\phi(Z;Y)} + \underbrace{\beta (\widehat{\mathcal{L}}_{\text{FR}}(\phi) + \widehat{\mathcal{L}}_{\text{JF}}(\phi))}_{\downarrow I_\phi(Z;X)}, \quad (19)$$

235 where $\beta \geq 0$ is the bottleneck multiplier. Minimizing $\mathbb{E}_{p(x,y)} \mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(y | z)]$ reduces
 236 $H(Y | Z)$ and increases $I_\phi(Z; Y)$ (since $H(Y)$ is fixed); the FR and JF terms jointly penalize
 237 $I_\phi(Z; X)$.
 238

239 **Natural Gradient on the Encoder.** Viewing $\{q_\phi(z | x)\}_\phi$ as a statistical manifold endowed with
 240 the Fisher–Rao metric, the natural gradient of any scalar objective $\mathcal{J}(\phi)$ is
 241

$$\tilde{\nabla}_\phi \mathcal{J} = F_\phi^{-1} \nabla_\phi \mathcal{J}, \quad F_\phi := \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} [\nabla_\phi \log q_\phi(z | x) \nabla_\phi \log q_\phi(z | x)^\top], \quad (20)$$

242 where the matrix F_ϕ in Eq. (20) is the Fisher–Rao metric tensor. We update the encoder by precon-
 243 ditioning the Euclidean gradient of the full objective:
 244

$$\phi_{t+1} = \phi_t - \eta_\phi F_{\phi_t}^{-1} \left(\nabla_\phi \mathbb{E}_{x,y,z} [-\log p_\theta(y | z)] + \beta [\nabla_\phi \mathcal{L}_{\text{FR}} + \nabla_\phi \mathcal{L}_{\text{JF}}] \right), \quad (21)$$

245 which is first-order invariant under smooth reparameterizations of ϕ and couples the distribution-
 246 level FR and geometry-level JF signals through a single preconditioner F_ϕ . We can also prove that
 247 the additive natural-gradient step $\phi^+ = \phi - \eta F_\phi^{-1} \nabla_\phi \mathcal{J}$ is a first-order approximation to the geodesic
 248 update in geometry.
 249

250 **Proposition 1** (Natural gradient equals the Riemannian gradient). *Let $\mathcal{M} = \{\phi : \phi \in \Theta \subset \mathbb{R}^d\}$
 251 be a regular statistical manifold endowed with the Fisher–Rao metric $g_\phi(u, v) := u^\top F_\phi v$, where
 252 $F_\phi = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} [\nabla_\phi \log q_\phi \nabla_\phi \log q_\phi^\top]$. For a scalar objective $\mathcal{J} : \mathcal{M} \rightarrow \mathbb{R}$, its Riemannian
 253 gradient at ϕ satisfies*

$$\text{grad } \mathcal{J}(\phi) = F_\phi^{-1} \nabla_\phi \mathcal{J}, \quad (22)$$

254 *i.e., the natural gradient $\tilde{\nabla}_\phi \mathcal{J} := F_\phi^{-1} \nabla_\phi \mathcal{J}$ is exactly the Riemannian gradient on (\mathcal{M}, g) .*
 255

256 See proof in Appendix E.
 257

258 **Proposition 2** (Steepest descent under the Fisher–Rao metric). *Let (\mathcal{M}, g) be endowed with the
 259 Fisher–Rao metric $g_\phi(u, v) = u^\top F_\phi v$ and let \mathcal{J} be smooth. The direction of steepest descent per
 260 unit FR length solves*

$$\min_{\|v\|_{g_\phi} \leq 1} \text{D}\mathcal{J}(\phi)[v], \quad (23)$$

261 *and the (unit-norm) optimizer is $v_* = -\frac{\text{grad } \mathcal{J}(\phi)}{\|\text{grad } \mathcal{J}(\phi)\|_{g_\phi}}$, with the convention $v_* = 0$ if $\text{grad } \mathcal{J}(\phi) =$
 262 0. In particular, by Proposition 1, its direction coincides with the negative natural gradient:
 263 $-\text{grad } \mathcal{J}(\phi) \equiv -F_\phi^{-1} \nabla_\phi \mathcal{J}$.*

270 *Proof sketch.* By the Riemannian gradient definition in (Amari & Nagaoka, 2000), $D\mathcal{J}(\phi)[v] =$
 271 $\langle \text{grad } \mathcal{J}(\phi), v \rangle_{g_\phi}$. Cauchy–Schwarz gives $\langle \text{grad } \mathcal{J}, v \rangle_{g_\phi} \geq -\|\text{grad } \mathcal{J}\|_{g_\phi} \|v\|_{g_\phi} \geq -\|\text{grad } \mathcal{J}\|_{g_\phi}$,
 272 with equality iff v is collinear with $-\text{grad } \mathcal{J}$ and $\|v\|_{g_\phi} = 1$. \square
 273

274 By Proposition 2, the steepest descent direction per unit FR length is $-\text{grad } \mathcal{J}(\phi)$. We thus update
 275 along this direction using the exponential map.

276 **Theorem 1** (Geodesic update via the exponential map). *Let $\text{Exp}_\phi : T_\phi \mathcal{M} \rightarrow \mathcal{M}$ be the Riemannian
 277 exponential map of the FR metric. The discrete update*

$$278 \quad \phi^+ = \text{Exp}_\phi(-\eta \text{grad } \mathcal{J}(\phi)) \\ 279$$

280 *lies on the unique FR geodesic γ starting at ϕ with initial velocity $\dot{\gamma}(0) = -\eta \text{grad } \mathcal{J}(\phi)$; i.e.,*
 281 $\phi^+ = \gamma(1)$.

282 See proof in Appendix F.

284 **Corollary 1** (First-order equivalence to the additive update). *Let \mathcal{R}_ϕ be any retraction on \mathcal{M} sat-
 285 isfying $\mathcal{R}_\phi(0) = \phi$ and $D\mathcal{R}_\phi(0) = \text{Id}$ (the exponential map is a canonical retraction). In local
 286 coordinates,*

$$287 \quad \text{Exp}_\phi(-\eta F_\phi^{-1} \nabla_\phi \mathcal{J}) = \phi - \eta F_\phi^{-1} \nabla_\phi \mathcal{J} + O(\eta^2).$$

288 *By Proposition 1, $\text{grad } \mathcal{J} = F_\phi^{-1} \nabla_\phi \mathcal{J}$. Hence the common additive natural-gradient step $\phi^+ =$
 289 $\phi - \eta F_\phi^{-1} \nabla_\phi \mathcal{J}$ is a first-order approximation to the geodesic update.*
 290

291 **Natural Gradient on the Decoder.** For the decoder, we use the natural gradient on $\{p_\theta(y | z)\}_\theta$:

$$293 \quad \theta_{t+1} = \theta_t - \eta_\theta F_\theta^{-1} \nabla_\theta \mathbb{E}_{x,y,z}[-\log p_\theta(y | z)], \quad (24)$$

294 where $F_\theta = \mathbb{E}_{q_\phi(z)} \mathbb{E}_{p_\theta(y|z)} [\nabla_\theta \log p_\theta(y | z) \nabla_\theta \log p_\theta(y | z)^\top]$. In practice, we compute $F_\theta^{-1} g$
 295 using scalable approximations and solvers, such as K-FAC (Martens & Grosse, 2015; Martens et al.,
 296 2018). As the page limitation, we present the whole G-IB algorithm in Appendix G.
 297

298 4 EXPERIMENTS

300 In this section, we conduct experiments to answer the following research questions (RQ) about G-IB:
 301

302 • **RQ1:** How does the proposed G-IB perform on information compression and prediction accuracy,
 303 as compared with the state-of-the-art IB solutions? (See Sections 4.2 and 4.3)
 304 • **RQ2:** How do different hyperparameters, such as the Lagrange multiplier β and representation
 305 dimensionality K , influence the G-IB? (See Sections 4.4 and 4.5)

307 4.1 EXPERIMENTAL SETTINGS

309 **Datasets.** We have conducted experiments on three widely adopted public datasets: MNIST (Deng,
 310 2012), CIFAR10 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2018), offering a range of ob-
 311 jective categories with varying levels of learning complexity. We present detailed statistics of all
 312 datasets and how do we use them in Appendix H.

313 **Models.** We select three model architectures of different sizes in our experiments: a 7-layer con-
 314 volutional neural network (CNN), a 5-layer multi-layer perceptron (MLP), and ResNet18. For the
 315 MNIST dataset, we employ two MLPs to form the G-IB model (one as the compression encoder and
 316 one as the task decoder). For CIFAR10 and CelebA, we employ the ResNet18 as the encoder and
 317 one MLP and CNN as decoder for CIFAR10 and CelebA respectively.

318 **Metrics.** We quantify model utility by top-1 classification **Accuracy** on the held-out test set. Infor-
 319 mation compression is assessed by the mutual information $I(Z;X)$ between the learned represen-
 320 tation and the input, estimated with MINE (Belghazi et al., 2018). To probe the leakage contained
 321 in Z , we perform two standard representation-level attacks: (i) a model inversion attack that re-
 322 constructs x from z (Fredrikson et al., 2015), evaluated by mean squared error (**MSE**; lower MSE
 323 indicates stronger leakage); and (ii) a membership inference attack (Shokri et al., 2017), evaluated
 by the membership inference accuracy (**MIA**) (higher values indicate stronger leakage).

324 Table 1: General Evaluation Results on image datasets, MNIST and CIFAR10, and CelebA. Results
 325 in bold are the best; those in italics are the second best.

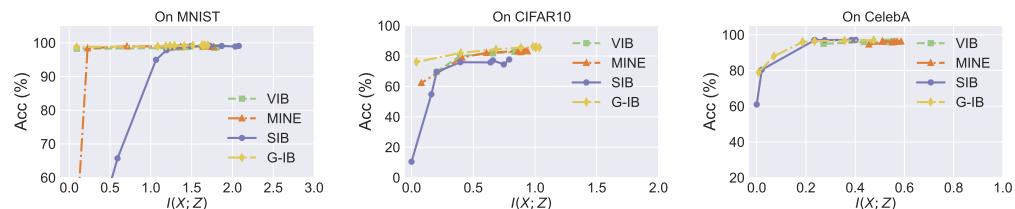
Methods	MNIST, $\beta = 0.0001, K = 128$			CIFAR10, $\beta = 0.0001, K = 128$			CelebA, $\beta = 0.0001, K = 128$		
	Accuracy	$I(X; Z)$	MSE	Accuracy	$I(X; Z)$	MSE	Accuracy	$I(X; Z)$	MSE
VIB (Alemi et al., 2016)	98.72%	1.81	0.034	82.65%	0.87	0.887	95.85%	0.55	0.054
SIB (Yang et al., 2025)	99.08%	1.86	0.037	75.81%	0.63	0.051	97.25%	0.38	0.072
MINE (Belghazi et al., 2018)	98.85%	1.76	0.032	82.64%	0.88	0.806	96.29%	0.51	0.048
AIB (Zhai & Zhang, 2022)	99.01%	1.69	0.043	82.63%	0.84	0.812	96.05%	0.52	0.053
G-IB (Our)	99.28%	1.69	0.043	85.54%	1.01	0.899	97.01%	0.47	0.066

326
 327
 328
 329
 330
 331
 332
 333 **Compared IB Benchmarks.** We compare G-IB against four representative Information Bottleneck
 334 (IB) variants: (1) the standard Variational IB (**VIB**) (Alemi et al., 2016); (2) an IB variant where
 335 the mutual information term is estimated with **MINE** (Belghazi et al., 2018); (3) the state-of-the-
 336 art Structured IB (**SIB**) focusing on structure-aware feature learning (Yang et al., 2025); and (4) the
 337 Adversarial IB (**AIB**) that incorporates adversarial regularization into the bottleneck (Zhai & Zhang,
 338 2022). For fairness, all methods use the same backbone, data preprocessing, and training schedule;
 339 hyperparameters are tuned on the validation set following the original papers where applicable.
 340

341 4.2 OVERALL EVALUATION OF G-IB

342
 343 **Setup.** We compare G-IB with four representative IB variants on MNIST, CIFAR10, and CelebA
 344 under the same backbone and schedule. We fix the Lagrange multiplier $\beta = 10^{-4}$ and the rep-
 345 resentation dimensionality $K = 128$ for an apples-to-apples comparison. We report top-1 accu-
 346 racy (higher is better), the estimated mutual information $I(X; Z)$ via MINE (lower $I(X; Z)$ value
 347 indicates stronger compression), and model-inversion MSE from reconstructing x from z (higher
 348 indicates less leakage) in Table 1.

349 **Results.** Across the three datasets, G-IB attains the best or second-best results on all metrics. On
 350 MNIST, G-IB achieves the highest accuracy and a (tied) lowest $I(X; Z)$, while matching the top
 351 inversion MSE. This suggests strong compression without sacrificing utility. On CIFAR10, G-IB
 352 delivers the best accuracy and the highest inversion MSE (least leakage). Although SIB attains the
 353 lowest $I(X; Z)$, G-IB offers a better accuracy–privacy trade-off overall. On CelebA, SIB slightly
 354 leads in accuracy and MI. G-IB ranks second with competitive accuracy and privacy (indicated by
 355 MSE), confirming robustness on a more fine-grained, structured dataset.

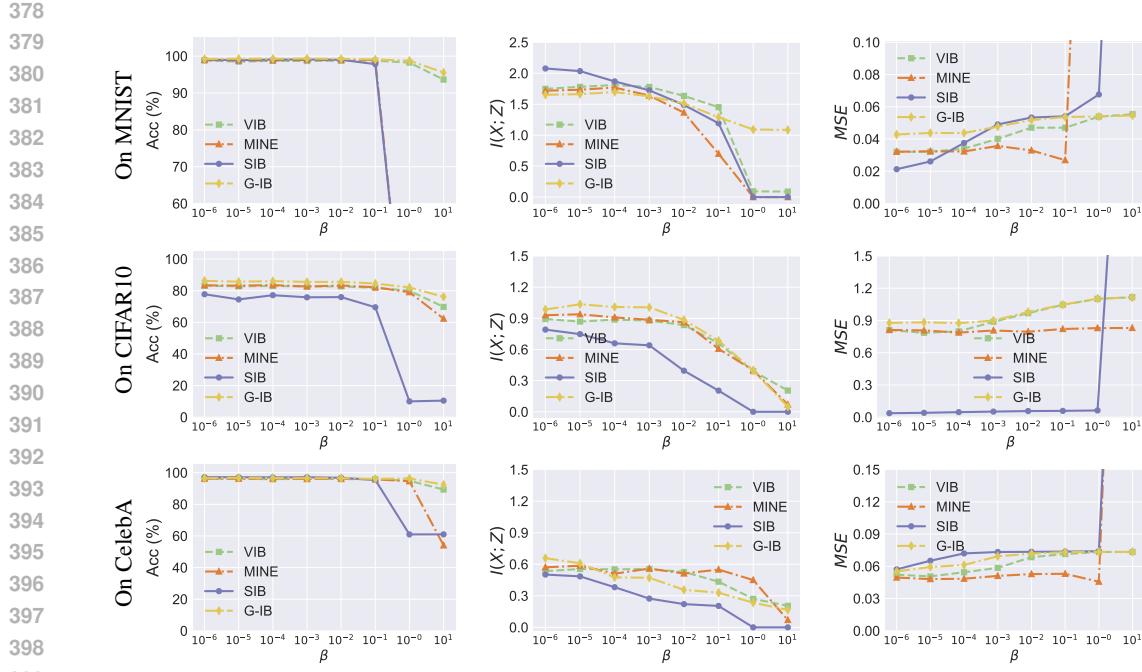


363 Figure 2: Evaluation of compression ratio and prediction accuracy from the information plane.
 364

365 4.3 INFORMATION-PLANE EVALUATION: COMPRESSION VS. ACCURACY

366
 367 **Setup.** We trace each method’s information plane on MNIST, CIFAR10, and CelebA by randomly
 368 selecting the Lagrange multiplier β from 10^{-6} to 10^1 . During this experiment, we fixed represen-
 369 tation size $K = 128$. For every β , we train to convergence and record test accuracy (higher is better)
 370 and the estimated mutual information $I(X; Z)$ via MINE as shown in Figure 2.

371 **Results.** In Figure 2, across datasets, G-IB’s curve is consistently shifted up and left relative to
 372 the baselines, achieving higher accuracy at a matched $I(X; Z)$, or a smaller $I(X; Z)$ at a matched
 373 accuracy (Pareto improvement). When compression is weak, i.e., large $I(X; Z)$, all methods reach
 374 a high-accuracy plateau. As compression strengthens, i.e., smaller $I(X; Z)$, VIB, MINE, and SIB
 375 exhibit clear accuracy drops, while G-IB maintains accuracy over a wider low- $I(X; Z)$ range before
 376 degrading. On CIFAR10 the gap is most visible, where G-IB preserves accuracy at lower $I(X; Z)$,
 377 And on CelebA, the curves cluster near the top but G-IB attains comparable accuracy with less
 378 information in Z .

Figure 3: Evaluation about the impact of the Bottleneck multiplier β .

4.4 ABLATION: INFLUENCE OF THE BOTTLENECK MULTIPLIER β

Setup. We study the impact of the Bottleneck multiplier by sweeping β on a logarithmic grid from 10^{-6} to 10^1 , with representation size fixed at $K = 128$. For each β , we train a model and report test accuracy (higher is better), the estimated mutual information $I(X; Z)$ (lower means stronger compression), and the model-inversion MSE from reconstructing x from z (higher means less leakage). Results for MNIST, CIFAR10, and CelebA are shown in Figure 3.

Results. (Left column) *Accuracy vs. β .* G-IB maintains the highest or near-highest accuracy over a wide range of β on all datasets. Baselines, especially MINE and SIB, exhibit sharp degradation when $\beta \geq 10^{-1}$; on CIFAR10 and CelebA, SIB collapses at large β .

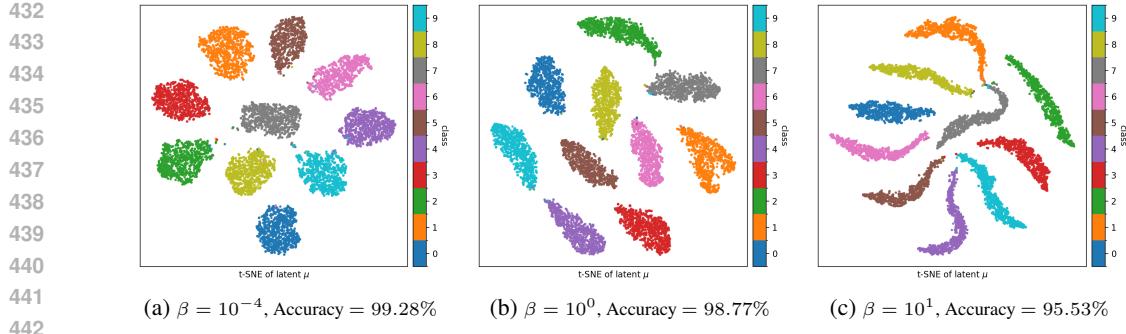
(Middle) *$I(X; Z)$ vs. β .* All methods show the expected monotonic decrease of $I(X; Z)$ as β increases. SIB achieves the smallest $I(X; Z)$ (strongest compression) but at the cost of the pronounced accuracy drop above, whereas G-IB attains competitive compression while preserving accuracy.

(Right) *MSE vs. β .* G-IB yields the best or second-best MSE across most β , indicating stronger resistance to inversion attacks. At extreme β ($10^0 - 10^1$), the MSE of some baselines spikes due to representation collapse. Reconstructions become effectively random, which inflates MSE but coincides with poor utility.

Qualitative visualization. To examine how the bottleneck strength shapes the representation, we visualize a 2-D t-SNE of the representation $\mu(x) = \mathbb{E}[Z | x]$ for 10,000 MNIST test images at three values of β (Figure 4); the CIFAR10 results are in Figure 6 in Appendix I. With a small Bottleneck multiplier ($\beta = 10^{-4}$), clusters are well separated and exhibit relatively large within-class spread, indicating that Z retains fine-grained input details. At $\beta = 10^0$, representation clusters contract toward class-wise prototypes while remaining separable. At $\beta = 10^1$, representation embeddings concentrate along narrow arcs near the class centers, consistent with stronger compression and the accuracy drop observed in Figure 3.

4.5 ABLATION: INFLUENCE OF THE REPRESENTATION DIMENSIONALITY K

Setup. We vary the representation dimensionality K from 2^1 to 2^9 with the Lagrange multiplier fixed at $\beta = 10^{-4}$. For each K , we report the accuracy, mutual information $I(X; Z)$, MSE, and MIA for different IB methods. Results for MNIST and CIFAR10 are shown in Figure 5.



443 Figure 4: Visualizing representation embeddings of posterior means $\mu(x)$ for 10,000 test images in
444 two dimensions on MNIST ($K = 128$). Colors denote true labels. From left to right: $\beta = 10^{-4}, 10^0$,
445 and 10^1 ; the corresponding test accuracies are shown below each panel. As β increases, within-class
446 dispersion shrinks and clusters move toward class-wise prototypes, indicating stronger compression;
447 accuracy decreases accordingly.

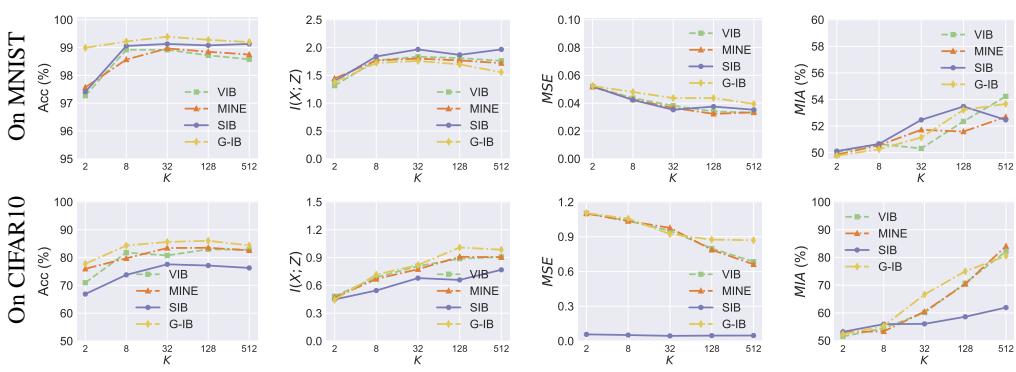


Figure 5: Evaluation about the impact of the Representation Dimensionality K .

462 **Results.** In the Left column of Figure 5, accuracy improves with K and then saturates (MNIST: 463 gains plateau around $K \geq 32$; CIFAR10: around $K \geq 128$). G-IB attains the highest or near- 464 highest accuracy across K , especially in the low-to-mid range. In the Mid-left column, as expected, 465 $I(X; Z)$ increases with K , indicating weaker compression when the latent space is wider. Tail 466 non-monotonicity on MNIST at very large K is minor and likely due to the representation dimen- 467 sionality of $K = 32$ is already large enough for MNIST dataset. The Mid-right column shows that 468 MSE generally decreases as K grows (reconstructions become easier), reflecting increased leakage 469 with higher-dimensional Z . G-IB keeps MSE competitively high (i.e., more resistant to inversion) 470 at small and medium K . In the Right column, MIA success rises with K on both datasets, corrobor- 471 ating that larger Z carries more membership signal. Across a broad range of K , G-IB remains 472 competitive; when K is small-to-medium, it attains a favorable utility–privacy balance.

5 SUMMARY AND FUTURE WORK

476 We propose the G-IB to solve the IB problem from a information geometry perspective. The key 477 ingredients are (i) a distribution-level Fisher–Rao (FR) discrepancy that locally matches KL to sec- 478 ond order and is invariant under smooth reparameterizations of the latent, and (ii) a geometry-level 479 Jacobian–Frobenius (JF) penalty that provides a local capacity-type upper bound on $I_\phi(Z; X)$ by 480 discouraging pullback volume expansion. On the optimization side, we propose a natural gradient 481 that aligned updates with the FR metric and prove that the standard additive natural-gradient step is 482 first-order equivalent to the exponential-map (geodesic) update.

483 Promising directions include tightening the FR-based proxies beyond the local (second-order) 484 regime and replacing the trace relaxation with sharper spectral approximations; connecting JF con- 485 trolled smoothness to robustness and generalization bounds; and extending G-IB to federated and 486 privacy-preserving settings where FR and JF controls might yield verifiable unlearning guarantees.

486 STATEMENTS
487

488 **Ethic Statements.** Our study only involves the use of images of public datasets, with a long history
489 of works having used these images for research. There are no new ethical implications for the
490 humans within the datasets in this paper. Our work does not contain any user participants, and the
491 outcome of this research influences benefits all individuals equally. As such “Respect for Persons”
492 is satisfied.

493 **Reproducibility Statements.** To ensure the reproducibility of our research, we are committed to
494 ensuring that our research is transparent, reproducible, and accessible to the broader community.
495 The source code and the artifact of the G-IB is available at <https://anonymous.4open.science/r/G-IB-0569>.
496
497

498 REFERENCES
499

500 Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations
501 through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.

503 Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information
504 bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

506 Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

508 Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American
509 Mathematical Soc., 2000.

510 Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron
511 Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine
512 learning*, pp. 531–540. PMLR, 2018.

513 Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A
514 contrastive log-ratio upper bound of mutual information. In *International conference on machine
515 learning*, pp. 1779–1788. PMLR, 2020.

517 Li Deng. The mnist database of handwritten digit images for machine learning research [best of the
518 web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

519 Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression
520 for lossless prediction. *Advances in Neural Information Processing Systems*, 34:14014–14028,
521 2021.

523 Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confi-
524 dence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC confer-
525 ence on computer and communications security*, pp. 1322–1333, 2015.

526 Jung-Ho Hong, Ho-Joong Kim, Kyu-Sung Jeon, and Seong-Whan Lee. Comprehensive information
527 bottleneck for unveiling universal attribution to interpret vision transformers. In *Proceedings of
528 the Computer Vision and Pattern Recognition Conference*, pp. 25166–25175, 2025.

530 Dou Hu, Lingwei Wei, Yixin Liu, Wei Zhou, and Songlin Hu. Structured probabilistic coding.
531 In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12491–12501,
532 2024.

533 Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

535 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.
536 2009.

537 Sun-Yuan Kung. A compressive privacy approach to generalized information bottleneck and privacy
538 funnel problems. *Journal of the Franklin Institute*, 355(4):1846–1872, 2018.

539 Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

540 Nunzio Alexandro Letizia, Nicola Novello, and Andrea M Tonello. Mutual information estimation
 541 via f -divergence and data derangements. *Advances in Neural Information Processing Systems*,
 542 37:105114–105150, 2024.

543 Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. In-
 544 variant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference*
 545 *on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022.

546 Hongru Li, Jiawei Shao, Hengtao He, Shenghui Song, Jun Zhang, and Khaled B Letaief. Tackling
 547 distribution shifts in task-oriented communication with information bottleneck. *IEEE Journal on*
 548 *Selected Areas in Communications*, 2025a.

549 Jin Li, Yaoming Wang, Xiaopeng Zhang, Dongsheng Jiang, Wenrui Dai, Chenglin Li, Hongkai
 550 Xiong, and Qi Tian. Contrastive learning via variational information bottleneck. *IEEE Transac-*
 551 *tions on Pattern Analysis and Machine Intelligence*, 2025b.

552 Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba)
 553 dataset. *Retrieved August*, 15(2018):11, 2018.

554 James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate
 555 curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

556 James Martens, Jimmy Ba, and Matt Johnson. Kronecker-factored curvature approximations for
 557 recurrent neural networks. In *International Conference on Learning Representations*, 2018.

558 David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information.
 559 In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

560 Behrooz Razeghi, Flavio P Calmon, Deniz Gunduz, and Slava Voloshynovskiy. Bottlenecks club:
 561 Unifying information-theoretic trade-offs among complexity, leakage, and utility. *IEEE Transac-*
 562 *tions on Information Forensics and Security*, 18:2060–2075, 2023.

563 Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcz, Robie Gonzales, Subhabrata Ma-
 564 jumdar, Hassan Sajjad, Frank Rudzicz, et al. Representation noising: A defence mechanism
 565 against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–
 566 12676, 2024.

567 Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of
 568 deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference*
 569 *on artificial intelligence*, volume 32, 2018.

570 Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference at-
 571 tacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*,
 572 pp. 3–18. IEEE, 2017.

573 Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via informa-
 574 tion. *arXiv preprint arXiv:1703.00810*, 2017.

575 Yang Song, Chong Xiao Wang, and Wee Peng Tay. Compressive privacy for a linear dynamical
 576 system. *IEEE Transactions on Information Forensics and Security*, 15:895–910, 2019.

577 Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and Philip S Yu. Graph
 578 structure learning with variational information bottleneck. In *Proceedings of the AAAI conference*
 579 *on artificial intelligence*, volume 36, pp. 4165–4174, 2022.

580 Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv*
 581 *preprint physics/0004057*, 2000.

582 Bo-Wei Tseng and Pei-Yuan Wu. Compressive privacy generative adversarial network. *IEEE Trans-*
 583 *actions on Information Forensics and Security*, 15:2499–2513, 2020.

584 Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. Multi-view information-bottleneck
 585 representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, vol-
 586 ume 35, pp. 10085–10092, 2021.

594 Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Scu: An efficient machine unlearning
 595 scheme for deep learning enabled semantic communications. *IEEE Transactions on Information*
 596 *Forensics and Security*, 2024.

597 Chunyu Wei, Jian Liang, Di Liu, and Fei Wang. Contrastive graph structure learning via information
 598 bottleneck for recommendation. *Advances in neural information processing systems*, 35:20407–
 600 20420, 2022.

601 Tailin Wu, Ian Fischer, Isaac L Chuang, and Max Tegmark. Learnability for the information bottle-
 602 neck. In *Uncertainty in Artificial Intelligence*, pp. 1050–1060. PMLR, 2020a.

603 Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. *Advances in*
 604 *Neural Information Processing Systems*, 33:20437–20448, 2020b.

605 Songjie Xie, Shuai Ma, Ming Ding, Yuanming Shi, Mingjian Tang, and Youlong Wu. Robust in-
 606 formation bottleneck for task-oriented communication with digital modulation. *IEEE Journal on*
 607 *Selected Areas in Communications*, 41(8):2577–2591, 2023.

608 Mengting Xu, Tao Zhang, Zhongnian Li, and Daoqiang Zhang. Infoat: Improving adversarial
 609 training using the information bottleneck principle. *IEEE Transactions on Neural Networks and*
 610 *Learning Systems*, 35(1):1255–1264, 2022.

611 Zhengze Xu, Dongyue Wu, Changqian Yu, Xiangxiang Chu, Nong Sang, and Changxin Gao. Sct-
 612 net: Single-branch cnn with transformer semantic information for real-time segmentation. In
 613 *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6378–6386, 2024.

614 Hanzhe Yang, Youlong Wu, Dingzhu Wen, Yong Zhou, and Yuanming Shi. Structured ib: Improving
 615 information bottleneck with structured feature learning. In *Proceedings of the AAAI Conference*
 616 *on Artificial Intelligence*, volume 39, pp. 21922–21928, 2025.

617 Junchi Yu, Jie Cao, and Ran He. Improving subgraph recognition with variational graph informa-
 618 tion bottleneck. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*
 619 *recognition*, pp. 19396–19405, 2022.

620 Shujian Yu, Xi Yu, Sigurd Løkse, Robert Jenssen, and Jose C Principe. Cauchy-schwarz divergence
 621 information bottleneck for regression. In *The Twelfth International Conference on Learning Rep-*
 622 *resentations*, 2024.

623 Penglong Zhai and Shihua Zhang. Adversarial information bottleneck. *IEEE Transactions on Neural*
 624 *Networks and Learning Systems*, 35(1):221–230, 2022.

625 Chenhan Zhang, Weiqi Wang, James JQ Yu, and Shui Yu. Extracting privacy-preserving subgraphs
 626 in federated graph learning using information bottleneck. In *Proceedings of the 2023 ACM Asia*
 627 *Conference on Computer and Communications Security*, pp. 109–121, 2023.

628 Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmen-
 629 tation for improved generalization and robustness. *Advances in Neural Information Processing*
 630 *Systems*, 33:14435–14447, 2020.

631

632 A LLM USAGE DECLARATION

633 The authors declare that Large Language Models (LLMs) were used for grammar correction and
 634 text refinement. All research ideas, analyses, results, tables, and figures presented in this paper are
 635 original contributions by the authors and were not generated by LLMs.

636 B RELATED WORK

637 The Information Bottleneck (IB) Lagrangian (Tishby et al., 2000) has been widely studied in rep-
 638 resentation learning (Achille & Soatto, 2018; Rosati et al., 2024) and practical training techniques
 639 (Xu et al., 2022; Li et al., 2025b). A practical deep implementation is the Variational Information

648 Bottleneck (VIB) (Alemi et al., 2016), which introduces a variational encoder $q_\phi(z \mid x)$, a prior
 649 $r(z)$, and a decoder/classifier $q_\theta(y \mid z)$, and uses the following bounds:
 650

$$\begin{aligned} 653 \quad I(Z; X) &= \mathbb{E}_{p(x)}[\text{KL}(q_\phi(z \mid x) \parallel q_\phi(z))] \leq \mathbb{E}_{p(x)}[\text{KL}(q_\phi(z \mid x) \parallel r(z))] \\ 654 &= \mathbb{E}_{p(x)q_\phi(z \mid x)} \log q_\phi(z \mid x) - \mathbb{E}_{q_\phi(z)} \log r(z), \end{aligned} \quad (25)$$

$$661 \quad I(Z; Y) = H(Y) - H(Y \mid Z) \geq H(Y) + \mathbb{E}_{p(x,y)q_\phi(z \mid x)} \log q_\theta(y \mid z). \quad (26)$$

665 Here $q_\phi(z) = \int q_\phi(z \mid x)p(x)dx$ is the encoder marginal. Eq. (25) uses $\text{KL}(q_\phi(z) \parallel r(z)) \geq 0$
 666 (with equality if $r(z) = q_\phi(z)$). Eq. (26) follows from the non-negativity of $\text{KL}(p(y \mid z) \parallel q_\theta(y \mid z))$
 667 (cross-entropy bound).

668 Beyond VIB, a complementary line of work replaces the prior–KL surrogate with neural mutual-
 669 information estimators. These methods train a critic to optimize variational bounds on MI and
 670 related quantities, enabling more direct optimization of the IB objective. Representative examples
 671 include MINE (Belghazi et al., 2018), which maximizes a Donsker–Varadhan lower bound; NWJ
 672 and related f -divergence bounds (Letizia et al., 2024); and CLUB, which provides an upper bound
 673 useful for penalizing $I(Z; X)$ (Cheng et al., 2020). These estimators remove the need to choose a
 674 prior $r(z)$, but introduce practical trade-offs (bias/variance, saturation for large MI, dependence on
 675 negatives and critic capacity), as analyzed by McAllester & Stratos (2020).

676 Moreover, IB methods are studies across diverse scenarios. In graph learning, IB appears as the
 677 Graph Information Bottleneck (GIB) for minimal-sufficient node and structure representations that
 678 are also robust to perturbations (Wu et al., 2020b; Sun et al., 2022); follow-ups include variational
 679 GIB for subgraph recognition and VIB-guided graph structure learning that jointly optimizes topol-
 680 ogy and features (Yu et al., 2022). In semantic communication, IB gives a principled way to trans-
 681 mit meaning rather than raw bits. Task-oriented links use variational IB to trade informativeness vs.
 682 channel robustness and to handle distribution shift (Xie et al., 2023; Li et al., 2025a). The IB method
 683 also has informed the design of contrastive objectives and augmentation strategies, encouraging rep-
 684 resentations that discard nuisance variation yet retain label-sufficient information (Wei et al., 2022;
 685 Li et al., 2025b; Zhao et al., 2020; Xu et al., 2022).

686 Another important application scenario of IB is compressive privacy for privacy preserving. It
 687 achieves privacy protection of original data when participating in machine learning service via com-
 688 pressing the original data into low-dimensional space for the target machine learning task (Kung,
 689 2018; Song et al., 2019; Zhang et al., 2023). Moreover, (Tseng & Wu, 2020) proposed a new
 690 privacy-preserving generative adversarial network (GAN) based on compressive pricacy. In their
 691 method (Tseng & Wu, 2020), the users will upload the compressed data Z to the server side to
 692 achieve a machine learning service.

693 While effective, most of the above approaches hinge on explicit or variational MI estimates in high
 694 dimensions, which can be brittle. We exploit the *geometry* of the statistical manifold to implement
 695 information bottleneck. Specifically, in information geometry, the Fisher–Rao metric endows dis-
 696 tributions with a Riemannian structure under which KL is locally the squared geodesic distance
 697 (Amari & Nagaoka, 2000); thus, a Fisher–Rao (FR) discrepancy $d_{\text{FR}}^2(q_\phi(z \mid x), r(z))$ provides a
 698 reparameterization-invariant surrogate for the $I(Z; X)$ compression term. Complementarily, view-
 699 ing the encoder mean map $\mu_\phi : \mathcal{X} \rightarrow \mathcal{Z}$ as inducing a pullback metric $J_\mu^\top J_\mu$ suggests penalizing
 700 local volume distortion, which connects to contractive and Jacobian-based regularization (Ross &
 701 Doshi-Velez, 2018). We present the detailed introduction of the geometric information bottleneck
 in the methodology section.

APPENDIX

C PROOF OF EQUATION 3

Proof. Using the log-factorization identity and Fubini/Tonelli

$$\begin{aligned}
 \text{KL}(p(x, z) \| q(x)r(z)) &= \iint p(x, z) \log \frac{p(x, z)}{q(x)r(z)} dx dz \\
 &= \iint p(x, z) \log \frac{p(x, z)}{p_X(x)p_Z(z)} dx dz \\
 &\quad + \iint p(x, z) \log \frac{p_X(x)}{q(x)} dx dz + \iint p(x, z) \log \frac{p_Z(z)}{r(z)} dx dz \\
 &= \underbrace{\text{KL}(p(x, z) \| p_X(x)p_Z(z))}_{= I(X; Z)} \\
 &\quad + \underbrace{\int p_X(x) \log \frac{p_X(x)}{q(x)} dx}_{= \text{KL}(p_X \| q)} + \underbrace{\int p_Z(z) \log \frac{p_Z(z)}{r(z)} dz}_{= \text{KL}(p_Z \| r)}.
 \end{aligned}$$

The last step uses Fubini's theorem to integrate out z and x respectively. Because each KL term is nonnegative, the minimum over q, r is achieved at $q = p_X, r = p_Z$, with value $I(X; Z)$. \square

Remark 1 (Information-geometric Pythagorean relation). *Since \mathcal{I} is flat in the appropriate dual affine coordinates, the above decomposition is also the IG “Pythagorean theorem”:*

$$\text{KL}(p \| qr) = \text{KL}(p \| p_X p_Z) + \text{KL}(p_X p_Z \| qr), \quad qr \in \mathcal{I}.$$

Thus $p_X p_Z$ is the e -projection of p onto \mathcal{I} , and $I(X; Z)$ is exactly the projection distance.

D PROOF OF LOCAL SECOND-ORDER (FR).

We show that for a regular parametric family $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ and θ' near θ , i.e., $\Delta = \theta' - \theta$,

$$\text{KL}(p_\theta \| p_{\theta'}) = \frac{1}{2} \Delta^\top F(\theta) \Delta + o(\|\Delta\|^2) = \frac{1}{2} d_{\text{FR}}(p_\theta, p_{\theta'})^2 + o(\|\Delta\|^2),$$

where $F(\theta)$ is the Fisher information and d_{FR} is the Fisher–Rao distance. Throughout we assume standard regularity: common support, differentiability up to second order in θ , finiteness of $F(\theta)$, and interchange of expectation and differentiation.

Write

$$\text{KL}(p_\theta \| p_{\theta'}) = \mathbb{E}_{p_\theta} [\log p_\theta(Z) - \log p_{\theta'}(Z)].$$

Fix z and expand $\log p_{\theta'}(z)$ at θ :

$$\log p_{\theta'}(z) = \log p_\theta(z) + \Delta^\top \nabla_\theta \log p_\theta(z) + \frac{1}{2} \Delta^\top \nabla_\theta^2 \log p_\theta(z) \Delta + o(\|\Delta\|^2).$$

Subtract from $\log p_\theta(z)$, take \mathbb{E}_{p_θ} , and use $\mathbb{E}_{p_\theta} [\nabla_\theta \log p_\theta(Z)] = 0$ (zero mean score) to obtain

$$\text{KL}(p_\theta \| p_{\theta'}) = -\frac{1}{2} \Delta^\top \mathbb{E}_{p_\theta} [\nabla_\theta^2 \log p_\theta(Z)] \Delta + o(\|\Delta\|^2).$$

By the information identity,

$$F(\theta) = \mathbb{E}_{p_\theta} [\nabla_\theta \log p_\theta(Z) \nabla_\theta \log p_\theta(Z)^\top] = -\mathbb{E}_{p_\theta} [\nabla_\theta^2 \log p_\theta(Z)],$$

hence

$$\text{KL}(p_\theta \| p_{\theta'}) = \frac{1}{2} \Delta^\top F(\theta) \Delta + o(\|\Delta\|^2). \quad (27)$$

The FR metric is the Riemannian metric on Θ given by $g_\theta(u, v) = u^\top F(\theta) v$ for tangent vectors $u, v \in \mathbb{R}^d$. Let $\gamma : [0, 1] \rightarrow \Theta$ be any C^1 curve with $\gamma(0) = \theta$ and $\gamma(1) = \theta'$. Its FR length is

$$L(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^\top F(\gamma(t)) \dot{\gamma}(t)} dt,$$

and the FR distance is $d_{\text{FR}}(p_\theta, p_{\theta'}) = \inf_\gamma L(\gamma)$. For $\|\Delta\| \rightarrow 0$, choose the straight segment $\gamma(t) = \theta + t\Delta$ to get

$$L(\gamma)^2 = \left(\int_0^1 \sqrt{\Delta^\top F(\theta + t\Delta) \Delta} dt \right)^2 = \Delta^\top F(\theta) \Delta + o(\|\Delta\|^2),$$

using continuity of F and a second-order expansion in t . Since the geodesic length is minimal,

$$d_{\text{FR}}(p_\theta, p_{\theta'})^2 = \Delta^\top F(\theta) \Delta + o(\|\Delta\|^2). \quad (28)$$

This follows from standard Riemannian geometry: in normal coordinates at θ , $d_{\text{FR}}(p_\theta, p_{\theta'})^2 = \Delta^\top F(\theta) \Delta + O(\|\Delta\|^3)$, hence equation 28. Combining equation 27 and equation 28 yields the local equivalence

$$\text{KL}(p_\theta \| p_{\theta'}) = \frac{1}{2} d_{\text{FR}}(p_\theta, p_{\theta'})^2 + o(\|\Delta\|^2).$$

Let $q_\phi(z|x)$ and a reference marginal $r(z)$ belong to $\{p_\theta\}$ with parameters $\theta(x)$ and θ_r respectively, and assume $\|\theta(x) - \theta_r\|$ is small for $p(x)$ -almost every x . Applying the pointwise result above with $\Delta(x) = \theta(x) - \theta_r$ gives

$$\text{KL}(q_\phi(z|x) \| r(z)) = \frac{1}{2} d_{\text{FR}}(q_\phi(z|x), r(z))^2 + o(\|\Delta(x)\|^2).$$

Taking $\mathbb{E}_{p(x)}$ and using dominated convergence (guaranteed by the regularity assumptions) yields

$$\mathbb{E}_{p(x)} \text{KL}(q_\phi(z|x) \| r(z)) = \frac{1}{2} \mathbb{E}_{p(x)} d_{\text{FR}}(q_\phi(z|x), r(z))^2 + o(\mathbb{E}_{p(x)} \|\Delta(x)\|^2).$$

When $r(z) = p_\phi(z)$ (the aggregate posterior), this is precisely the local second-order approximation of the compression term $\mathbb{E}_{p(x)} \text{KL}(q_\phi(z|x) \| p_\phi(z))$. \square

E PROOF OF PROPOSITION 1

Proof of Proposition 1. By definition, the Riemannian gradient $\text{grad } \mathcal{J}(\phi) \in T_\phi \mathcal{M} \simeq \mathbb{R}^d$ is the unique vector field satisfying, for all tangent directions $v \in T_\phi \mathcal{M}$,

$$\langle \text{grad } \mathcal{J}(\phi), v \rangle_{g_\phi} = D\mathcal{J}_\phi[v].$$

Under the Fisher–Rao metric, the inner product is $\langle u, v \rangle_{g_\phi} = u^\top F_\phi v$. In coordinates, the differential equals the Euclidean pairing with the usual gradient: $D\mathcal{J}_\phi[v] = v^\top \nabla_\phi \mathcal{J}$. Hence, for all v ,

$$v^\top F_\phi \text{grad } \mathcal{J}(\phi) = v^\top \nabla_\phi \mathcal{J}.$$

Since F_ϕ is positive definite at regular points, we conclude $F_\phi \text{grad } \mathcal{J}(\phi) = \nabla_\phi \mathcal{J}$, i.e. $\text{grad } \mathcal{J}(\phi) = F_\phi^{-1} \nabla_\phi \mathcal{J}$. \square

F PROOF OF THEOREM 1

Proof of Theorem 1. Let $v := -\eta \text{grad } \mathcal{J}(\phi) \in T_\phi \mathcal{M}$. By existence and uniqueness for the geodesic equation with the Levi–Civita connection of the FR metric, there exists $\varepsilon > 0$ and a unique geodesic $\gamma_v : (-\varepsilon, \varepsilon) \rightarrow \mathcal{M}$ such that $\gamma_v(0) = \phi$ and $\dot{\gamma}_v(0) = v$.

The Riemannian exponential map at ϕ is defined by

$$\text{Exp}_\phi(w) = \gamma_w(1) \quad \text{whenever } 1 \text{ lies in the domain of } \gamma_w,$$

equivalently $\text{Exp}_\phi(tw) = \gamma_w(t)$ for t in a neighborhood of 0. (For global well-definedness one may restrict to $\|w\|$ below the injectivity radius.) Applying this with $w = v$ gives

$$\phi^+ = \text{Exp}_\phi(v) = \gamma_v(1),$$

i.e., ϕ^+ lies on the unique FR geodesic starting at ϕ with initial velocity $\dot{\gamma}(0) = v = -\eta \text{grad } \mathcal{J}(\phi)$. \square

810

811

Algorithm 1: G-IB Natural-Gradient Step (per iteration)

812 **Input:** Current params (ϕ_t, θ_t) ; minibatch size B ; step sizes (η_ϕ, η_θ) ; bottleneck β ; # Hutchinson probes
 813 S ; damping λ ; Fisher approx. mode K-FAC

Output: Updated params $(\phi_{t+1}, \theta_{t+1})$

```

815 1 procedure GIB-Step ( $\phi_t, \theta_t, B, \eta_\phi, \eta_\theta, \beta, S, \lambda, mode$ ) :
816 2   // 1) Sample minibatch and latent codes
817 3   Draw minibatch  $\{(x_i, y_i)\}_{i=1}^B$ ;;
818 4    $z_i \sim q_{\phi_t}(z | x_i)$  by reparameterization;
819 5   // 2) Estimate FR/JF terms (Hutchinson + JVPs)
820 6   Estimate  $\widehat{\mathcal{L}}_{\text{FR}}$  and  $\widehat{\mathcal{L}}_{\text{JF}}$  using  $S$  probe vectors and JVPs;
821 7   // 3) Compute Euclidean gradients
822 8    $g_\theta \leftarrow \nabla_\theta \frac{1}{B} \sum_{i=1}^B [-\log p_{\theta_t}(y_i | z_i)];$ 
823 9    $g_\phi \leftarrow \nabla_\phi \left\{ \frac{1}{B} \sum_{i=1}^B [-\log p_{\theta_t}(y_i | z_i)] + \beta(\widehat{\mathcal{L}}_{\text{FR}} + \widehat{\mathcal{L}}_{\text{JF}}) \right\};$ 
824 10  // 4) Build Fisher approximations (Empirical Fisher or K-FAC)
825 11  if  $mode = K\text{-FAC}$  then
826 12   | Build layerwise Kronecker factors for encoder/decoder to obtain  $\widehat{F}_\phi$  and  $\widehat{F}_\theta$ ;
827 13    $\widehat{F}_\theta^\lambda \leftarrow \widehat{F}_\theta + \lambda I, \quad \widehat{F}_\phi^\lambda \leftarrow \widehat{F}_\phi + \lambda I;$ 
828 14   // 5) Solve for natural directions (no explicit inversion)
829 15   Find  $v_\theta$  s.t.  $(\widehat{F}_\theta^\lambda) v_\theta = g_\theta$  via Conjugate Gradient (CG);
830 16   Find  $v_\phi$  s.t.  $(\widehat{F}_\phi^\lambda) v_\phi = g_\phi$  via CG;
831 17   // 6) Parameter updates (natural gradients)
832 18    $\theta_{t+1} \leftarrow \theta_t - \eta_\theta v_\theta$  (Eq. (24));
833 19    $\phi_{t+1} \leftarrow \phi_t - \eta_\phi v_\phi$  (Eq. (21));
834 20   return  $(\phi_{t+1}, \theta_{t+1})$ ;

```

838

G THE G-JB ALGORITHM

840

We provide the algorithm of G-JB as the following Algorithm 1.

841

Algorithm 1 executes one G-IB training iteration with *natural-gradient* updates for the decoder θ and encoder ϕ under a K-FAC curvature mode. (Step 1) Given a minibatch $\{(x_i, y_i)\}_{i=1}^B$, latent codes are drawn via the reparameterized posterior $z_i \sim q_{\phi_t}(z \mid x_i)$. (Step 2) The geometry-aware bottleneck surrogates are estimated without forming full Jacobians: the Fisher–Rao proxy $\hat{\mathcal{L}}_{\text{FR}}$ and the Jacobian–Frobenius penalty $\hat{\mathcal{L}}_{\text{JF}}$ are computed using S Hutchinson probe vectors together with Jacobian–vector products (JVPs). (Step 3) We then compute the Euclidean gradients of the decoder NLL and the full encoder objective,

849
850
851

$$g_\theta = \nabla_\theta \frac{1}{B} \sum_{i=1}^B \left[-\log p_{\theta_t}(y_i \mid z_i) \right], \quad g_\phi = \nabla_\phi \left\{ \frac{1}{B} \sum_{i=1}^B \left[-\log p_{\theta_t}(y_i \mid z_i) \right] + \beta (\hat{\mathcal{L}}_{\text{FR}} + \hat{\mathcal{L}}_{\text{JF}}) \right\}.$$

852

(Step 4) To obtain natural directions, we build *layerwise* Kronecker-factored Fisher approximations for both networks. For each layer ℓ with weight matrix W_ℓ , K-FAC uses the block-diagonal model $F_\ell \approx A_\ell \otimes G_\ell$, where $A_\ell := \frac{1}{B} \sum_i a_{\ell,i} a_{\ell,i}^\top$ is the covariance of layer inputs (activations) and $G_\ell := \frac{1}{B} \sum_i g_{\ell,i} g_{\ell,i}^\top$ is the covariance of backpropagated pre-activation gradients. Tikhonov damping yields $\hat{F}_\theta^\lambda = \hat{F}_\theta + \lambda I$ and $\hat{F}_\phi^\lambda = \hat{F}_\phi + \lambda I$. (Step 5) Rather than inverting these matrices, we solve the linear systems $(\hat{F}_\theta^\lambda)v_\theta = g_\theta$ and $(\hat{F}_\phi^\lambda)v_\phi = g_\phi$ via conjugate gradients (CG). Each CG iteration only needs Fisher–vector products, which K-FAC supplies efficiently: if V_ℓ reshapes the vector v to the layer’s weight shape, then

860

$$\text{FVP}_\ell(v) = \text{vec}\left((G_\ell + \lambda I) V_\ell (A_\ell + \lambda I)\right).$$

862

(Step 6) Finally, parameters are updated along the natural directions: $\theta_{t+1} = \theta_t - \eta_\theta v_\theta$ (cf. Eq. (24)) and $\phi_{t+1} = \phi_t - \eta_\phi v_\phi$ (cf. Eq. (21)). The hyperparameters $(p_+, n_0, \beta, S, \lambda)$ control step sizes

864 compression strength, estimator variance, and conditioning, while K-FAC trades a faithful curvature
 865 signal for scalable, inversion-free natural-gradient steps.
 866

867 Table 2: Dataset statistics.
 868

869	Dataset	Feature Dimension	#. Classes	#. Samples
871	MNIST (Deng, 2012)	28×28×1	10	70,000
872	CIFAR10 (Krizhevsky et al., 2009)	32×32×3	10	60,000
873	CelebA (Liu et al., 2018)	178×218×3	2 (Gender)	202,599

875

H DATASETS

 876

877 The statistics of all datasets used in our experiments are listed in Table 2. Both MNIST and CIFAR10
 878 are used to train 10-class classification models. The experiment on CelebA is to identify the gender
 879 attributes of the face images. The task is a binary classification problem, different from the ones on
 880 MNIST and CIFAR10. These datasets offer a range of objective categories with varying levels of
 881 learning complexity. We also introduce them as below.
 882

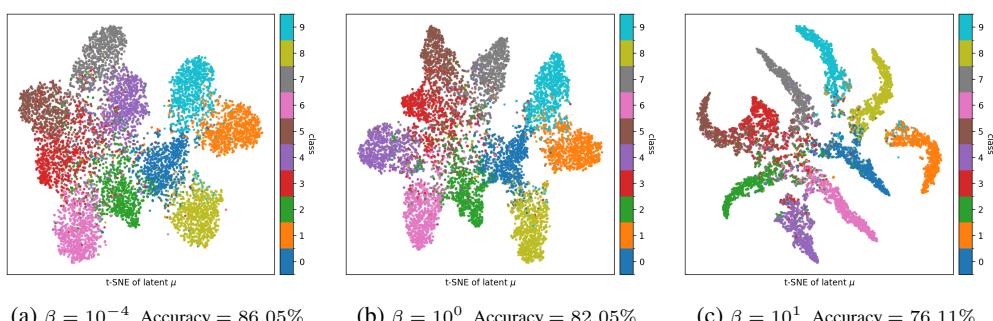
- 883 • **MNIST (Deng, 2012).** MNIST contains 60,000 handwritten digit images for the training and
 884 10,000 handwritten digit images for the testing. All these black and white digits are size normal-
 885 ized, and centered in a fixed-size image with 28×28 pixels.
- 886 • **CIFAR10 (Krizhevsky et al., 2009).** CIFAR10 dataset consists of 60,000 32x32 colour images in
 887 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.
- 888 • **CelebA (Liu et al., 2018).** CelebA is a large-scale face attributes dataset with more than 200,000
 889 celebrity images, each with 40 attribute annotations.

890

I ADDITIONAL EXPERIMENTS

 891

892 **Qualitative visualization.** To examine how the bottleneck strength shapes the representation, we
 893 visualize a 2-D t-SNE of the representation $\mu(x) = \mathbb{E}[Z \mid x]$ for 10,000 CIFAR10 test images
 894 at three values of β (Figure 6). With a small Bottleneck multiplier ($\beta = 10^{-4}$), clusters are well
 895 separated and exhibit relatively large within-class spread, indicating that Z retains fine-grained input
 896 details. At $\beta = 10^0$, representation clusters contract toward class-wise prototypes while remaining
 897 separable. At $\beta = 10^1$, representation embeddings concentrate along narrow arcs near the class
 898 centers, consistent with stronger compression and the accuracy drop observed in Figure 3.
 899



912 Figure 6: Visualizing representation embeddings of 10000 test images in two dimensions on CI-
 913 FAR10. The images are colored according to their true class label. As β becomes larger, we forget
 914 more about the input and the representation embedding of each class is compressed close to the
 915 average μ . We also report the test accuracy, which decreases as β increases.
 916
 917