Supplementary Material for Wasserstein Distributional Normalization

Sung Woo Park [1] Junseok Kwon [1]

1. Open-source Dataset

Transition matrix for CIFAR-10/100. For the experiment summarized in Table 1, we implemented open-source code to generate the noise transition matrix discussed by (Han et al., 2018), as well as the 9-layered CNN architecture (`https://github.com/bhanML/Co-teaching`).

Open-set noise. For the experiment summarized in Table 2, we used the same dataset for open-set noisy labels presented by (Lee et al., 2019) (`https://github.com/pokaxpoka/RoGNoisyLabel`).

Clothing1M. For the experiment summarized in Table 3, we used the open-source dataset presented by (Xiao et al., 2015) (`https://github.com/Cysu/noisy_label`).

2. Collaboration with Other Methods

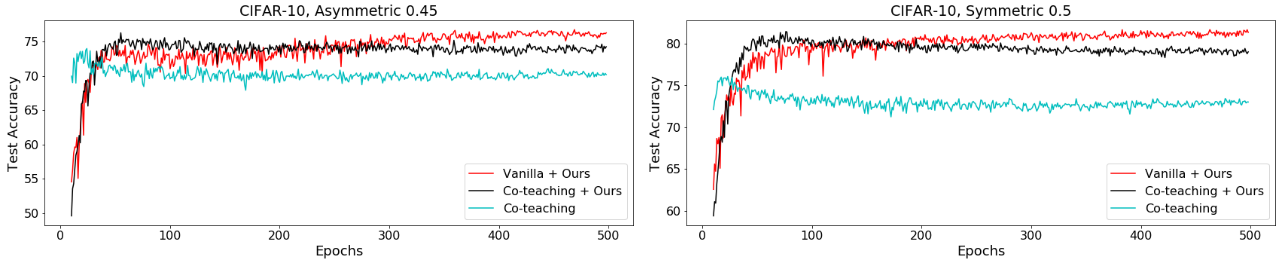

Figure 1. **Test accuracy for the proposed collaboration model with co-teaching.**

Our collaboration model with co-teaching achieved the most accurate performance for the CIFAR-100 dataset with asymmetric noise, which verifies that our WDN can be integrated into existing methods to improve their performance significantly, especially when the density of pre-logits is highly-concentrated. Fig.1 reveals that co-teaching quickly falls into over-parameterization and induces drastic drop in accuracy after the 15th-epoch. WDN$_{cot}$ also exhibited a slight accuracy drop. However, it surpassed the baseline co-teaching method by a large margin ($+7\%$) during training. This demonstrates that our enhanced samples X_T can alleviate the over-parameterization issues faced by conventional co-teaching models, which helps improve their accuracy significantly.

3. Comparisons to Related Works

Table 1 indicates that no previous methodology can conceptually include our method.

Because the solution to the Fokker-plank equation can be explicitly calculated without any additional parameter, our method is fully non-parametric (in terms of additional parameters against those required by the original neural network). By contrast, co-teaching is parametric, because it requires a clone network with additional parameters that are copies of those in the original network. Similarly, MLNT requires an additional teacher network for training, which also contains a large number of parameters.

Many methods based on small-loss criteria select certain samples, whereas our method uses the combination of ρN certain and $(1-\rho)N$ normalized uncertain samples. Therefore, our method can make the full use of the batches of training datasets, where $(1-\rho)N + \rho N = N$. Additionally, our method does not assume any class-dependent prior knowledge. Rather than considering class-wise prior knowledge, our method uses holistic information from both certain and uncertain samples (*i.e.*, Y and X_T) in the logit space. Other meta-class-based model, such as MLNT, assume class-wise meta prior knowledge

Table 1. **Differences of the proposed method for other methods.**

Methodology	Parametric	Class-dependency	Distillation	Sample-weight	Sample-selection
DivideMix	✓	✗	✗	✗	✓
Co-teaching	✓	✗	✓	✗	✓
JoCoR	✓	✗	✓	✗	✓
MLNT	✓	✓	✓	✗	✗
(Ren et al., 2018)	✗	✗	✗	✓	✗
NPCL	✗	✗	✗	✓	✗
GCE	✗	✗	✗	✓	✗
WDN	✗	✗	✗	✗	✗

from a teacher network.

In (Arazo et al., 2019), they assumed the beta-mixture model as a label distribution on the label space. However, due to the non-deterministic type of noisy label distribution, it sometimes fails to train with an extremely non-uniform type of noise. For example, (Arazo et al., 2019) reported failure cases with the Clothing1M dataset. It shows that fundamental assumption on noise model of *mixup* needs to be improved as a future work. Similar to this method, our work have trouble when dealing with synthetic asymmetric noise with a high ratio where a relatively large performance drop is observed (despite our method produces the second best performance in the table).

In most recent work (Li et al., 2019), they also adopted Co-train by implementing additional dual networks, which requires a very sophisticated methodology called Co-divide/guessing based on SSL. In their method, the Wasserstein distance between labeled and unlabeled probability measures is well-controlled. We believe that applying the OT/Markov theory (as in our paper) to their method will broaden understanding of the LNL problem.

In contrast to sample weight methods such as GCE and NPCL, which require prior knowledge regarding the cardinality of the training samples to be weighted, our method is free from this assumption, because our Wasserstein normalization can be applied in a batch-wise manner.

4. Technical Difficulty in Applying General Optimal Transport/Markov Theory to Label Space

Let X and Y be uncertain and certain samples in pre-softmax feature space, respectively. We consider the distributional constraint on the label-space (the space of $\sigma(X), \sigma(Y)$, where σ denotes the soft-max function). This space cannot appropriately define the objective function such as (7) of the main paper. Because all the samples in this label space is of the form $\sigma(X) = [a_1, a_2, \cdots, a_n]$ such that $\sum_{i=1}^{d} a_i = 1$, label-space is d-dimensional affine-simplex U_d, which is subset of the Euclidean space $U_d \subset \mathbb{R}^d$. In this case, the definition of Wasserstein space in (6) of the main paper is unacceptable while d_E is not a true metric on U_d. The Wasserstein space $\mathcal{P}_2(U_d)$ is merely investigated in the mathematical literature, which makes us unable to use all the technical details and assumptions, theories developed in the $\mathcal{P}_2(\mathbb{R}^d)$, which are theoretical ground of our work. However, if we interpret this problem slightly different point of view and consider pre-softmax $\mathbb{R}^d, \mathcal{P}_2(\mathbb{R}^d)$ as our base space, all the technical issues/problems when we try to use OT tools in $\mathcal{P}_2(U_d)$ can be overcome/ignored. while softmax is the non-parametric one-to-one function connecting pre-softmax feature space $\mathbb{R}^d$ to U_d, there exists a unique label in U_d as a mapped point of the manipulated uncertain samples. Even though our objects are defined on the pre-softmax space, the theoretical analysis in Proposition 3 of the main paper contains the softmax function to evaluate the concentration inequality of proposed transformation $\mathcal{F}$ affecting in label-space U_d.

5. Mathematical Background

In this section, we introduce important definitions, notations, and propositions used in our proofs and the main paper.

5.1. Notation

We denote $f_{\#}\mu$ as a push-forward of μ through f. $C_0^\infty(\mathbb{R}^d)$ denotes the set of ∞-class functions with compact support in $\mathbb{R}^d$. For the L_p-norm of the function f, we denote $\|f\|_{p,\xi} = (\int |f|^p d\xi)^{\frac{1}{p}}$. The Hessian matrix of the function f is denoted as $\mathbf{Hess}[f] = [\partial_i \partial_j f]_{i,j}^d$. $\mathbf{Sym}_d^+$ denotes the space for semi-definite positive symmetric matrices of size $d \times d$. $\|f\|_{Lip}$ denotes the Lipschitz norm of the function f. For any matrix $A \in \mathbb{M}_d$, we let $\|A\|_{op}$ denote the operator norm of A.

5.2. Diffusion-invariance and Hyper-contractivity

Definition 1. *The Markov semigroup $(P_t)_{t\geq 0}$ in $\mathbb{R}^d$ acting on a function $f \in C_0^\infty$ is defined as follows:*

$$P_t f(x) = \int f(x') p_t(x, dx'), \tag{1}$$

where $p_t(x, dx')$ is a transition kernel that is the probability measure for all $t \geq 0$.

Definition 2. ***(Diffusion Operator)*** *Given a Markov semi-group P_t at time t, the diffusion operator (i.e., infinitesimal generator) $\mathcal{L}$ of P_t is defined as*

$$\mathcal{L}g(y) = \lim_{t\to 0} \frac{1}{t}\left(P_t g(y) - g(y)\right) = \sum_{i,j} \frac{\partial^2}{\partial y_i \partial y_j} B^{ij}(y) g(y) - \sum_i A^i(y) \frac{\partial}{\partial y_i} g(y), \tag{2}$$

where B and A are matrix and vector-valued measurable functions, respectively. B^{ij} denotes the (i,j)-th function of B and A^i denotes the i-th component function of A.

Definition 3. *(Diffusion-invariant Measure) Given the diffusion operator $\mathcal{L}$, the probability measure μ is considered to be invariant measure to $\mathcal{L}$ when $\mathbb{E}_{X\sim\mu}[\mathcal{L}f(X)] = 0$ for any $f \in C_0^\infty$.*

Lemma 1. ***(Infinitesimal generator for the multivariate Gaussian measure, (Bolley & Gentil, 2010).)*** *The Gaussian measure $\mathcal{N}_\xi := \mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi)$ with a mean $\mathbf{m}_\xi$ and covariance Σ_ξ is an invariant measure according to the following diffusion-operator $\mathcal{L}$:*

$$\mathcal{L}f(x) = \Sigma_\xi \mathbf{Hess}[f](x) - (x - \mathbf{m}_\xi)^T \nabla f(x), \quad \forall f \in C_0^\infty(\mathbb{R}^d), \tag{3}$$

where $B^{ij}(x) := [\Sigma_\xi]_{ij}$ is a constant function, and $A^i(x) := x^i - \mathbf{m}_\xi^i$.

This generator serves as our main tool for the geometric analysis of the upper bound ε. In Section 5 in the main paper, we introduced an approximate upper-bound $\hat{K}_2(\mu)$ without any general description of the inequality involved. We now introduce the underlying mathematics for (33). Because our detour measure is Gaussian, there is a unique semi-group $P_t h$ called the *multidimensional Ornstein-Ulenbeck semi-group* that is invariant to $\mathcal{N}_\xi$. Specifically, P_t is defined as follows:

$$P_s h(X) = \mathbb{E}_{Z\sim\mathcal{N}_\mathbf{I}}\left[h\left(e^{-s}X + \sqrt{1-e^{-2s}}(\Sigma_\xi^{\frac{1}{2}} Z + \mathbf{m}_\xi)\right)\right], \quad \forall h \in C_0^\infty. \tag{4}$$

The invariance property of P_t relative to our detour measure is naturally induced by the following Proposition:

Proposition 1. *We define $C : \mathbb{R}^d \to \mathbb{R}^d$ and $C(X) = \mathbf{A}X + \mathbf{b}$ such that $\mathbf{A} \in \mathbf{Sym}_d^+, \mathbf{b} \in \mathbb{R}^d$, and select an arbitrary smooth $h \in C_0^\infty(\mathbb{R}^d)$. We then define the diffusion Markov semi-group $P_s h$ as follows:*

$$P_s h(X) = \mathbb{E}_{Z\sim\mathcal{N}}\left[h\left(e^{-s}X + \sqrt{1-e^{-2s}}C(Z)\right)\right]. \tag{5}$$

Then, $\mathcal{N}(\mathbf{A^2}, \mathbf{b})$ is invariant with respect to P_s, meaning the following equality holds for every h and $s \geq 0$:

$$\int_{\mathbb{R}^d} [P_s h(X) - h(X)] d\mathcal{N}(\mathbf{A^2}, \mathbf{b})(X) = 0. \tag{6}$$

Proof. For simplicity, we denote $\mathcal{N}(\mathbf{A^2}, \mathbf{b}) := \mathcal{N}_C$.

$$\begin{aligned} \int P_s h(X) d\mathcal{N}_C(X) &= \int\int h(e^{-s}X + \sqrt{1-e^{-2s}}C(Z)) d\mathcal{N}_C(X) d\mathcal{N}(Z) \\ &= \int\int h \circ C(e^{-s}Z' + \sqrt{1-e^{-2s}}Z) d\mathcal{N}(Z') d\mathcal{N}(Z). \end{aligned} \tag{7}$$

The second equality holds because C is linear in $\mathbb{R}^d$. Let $e^{-s} = \cos\theta$ and $e^{-2s} = \sin\theta$ for any $0 \leq \theta \leq 2\pi$. Then, we define ϕ as $\phi(Z', Z) = e^{-s}Z' + \sqrt{1-e^{-2s}}Z = \cos(\theta)Z' + \sin(\theta)Z$, and $\pi(Z', Z) = Z$. Based on the rotation property of the standard Gaussian measure, one can induce the following equality.

$$(\mathcal{N} \otimes \mathcal{N}) \circ (C \circ \phi)^{-1} = ((\mathcal{N} \otimes \mathcal{N}) \circ \phi^{-1}) \circ C^{-1} = \mathcal{N} \circ C^{-1}. \tag{8}$$

However, we know that $d\mathcal{N}[C^{-1}(X)] = d\mathcal{N}_C(X) = \left((2\pi)^d |\mathbf{A}^2|\right)^{-\frac{1}{2}} e^{-0.5(X-\mathbf{b})^T \mathbf{A}^{-2}(X-\mathbf{b})}$. By combining (7) and (8), one can derive the following result:

$$\begin{aligned}\int h \circ C(e^{-s}Z' + \sqrt{1-e^{-2s}}Z) d[\mathcal{N} \otimes \mathcal{N}] &= \int h(X) d\left[(\mathcal{N} \otimes \mathcal{N}) \circ \phi^{-1} \circ C^{-1}\right](X) \\ &= \int h(X) d[\mathcal{N} \circ C^{-1}](X) = \int h(X) d\mathcal{N}[C^{-1}(X)] \\ &= \int h(X) d\mathcal{N}_C(X).\end{aligned} \tag{9}$$

□

Proposition 1 demonstrates the invariance property of the defined semi-group. If we set $A = \Sigma_\xi^{\frac{1}{2}}, \mathbf{b} = \mathbf{m}_\xi$, then we can recover (4).

We are now ready to define the approximation of $K_2(\mu)$ in terms of semi-group invariance. Specifically, for any real-valued smooth h, we define the following inequality:

$$\begin{aligned}\hat{K}_2(\mu) &= \mathbb{E}_{X\sim\mu}[\mathcal{L}h(X)] = \lim_{s\to 0} \mathbb{E}_{X\sim\mu}\left[\frac{1}{s}(P_s h(X) - h(X))\right] \\ &= \lim_{s\to 0} \frac{1}{s}\mathbb{E}_{X,Z\sim\mathcal{N}_\mathbf{I}}\left[h\left(e^{-s}X + \sqrt{1-e^{-2s}}(\Sigma_\xi^{\frac{1}{2}}Z + \mathbf{m}_\xi)\right) - h(X)\right] \leq K_2(\mu).\end{aligned} \tag{10}$$

This inequality holds if h is selected to induce a supremum over the set C_0^∞, where $\sup_h \hat{K}_2(\mu, h) = \sup_h \mathbb{E}_{X\sim\mu}[\mathcal{L}h(X)] = K_2(\mu)$. Although a more sophisticated design for the test function h will induce a tighter upper bound for $\hat{K}_2$, we determined that the L_2-norm is generally sufficient.

Definition 4. ***(Diffuseness of the probability measure)*** *We define the integral operator $K_2 : \mathcal{W}_2(\mathbb{R}^d) \to \mathbb{R}^+$ as follows:*

$$K_2(\mu) = \sqrt{\sup_{f\in C_0^\infty} \int_{\mathbb{R}^d} |\mathcal{L}f(x)|\, d\mu(x)}. \tag{11}$$

According to Definition 3, we know that $\int \mathcal{L}f(X)d\mathcal{N}_\xi(X) = 0$ for any f. Based on this observation, it is intuitive that K_2 estimates how the probability measure ξ is distorted in terms of diffusion invariance. While this measure takes a supremum over the function space C_0^∞, it searches for a function that enables the estimation of maximal distortion. Because the value of K_2 is entirely dependent on the structure of μ, K_2 can be considered as a constant for the sake of simplicity if the uncertain measure μ is fixed over one iteration of training.

Definition 5. ***(Diffusion carré du champ)*** *Let $f, g \in C_0^\infty(\mathbb{R}^d)$. Then, we define a bilinear form Γ_c in $C_0^\infty(\mathbb{R}^d) \times C_0^\infty(\mathbb{R}^d)$ as*

$$\Gamma_e(f,g) = \frac{1}{2}[\mathcal{L}\Gamma_{e-1}(fg) - \Gamma_{e-1}(f\mathcal{L}g) - \Gamma_{e-1}(g\mathcal{L}f)], \quad e \geq 1. \tag{12}$$

We also denote $\Gamma(f) \equiv \Gamma(f,f)$. The bilinear form Γ can be considered as a generalization of the integration by the parts formula, where $\int f\mathcal{L}g + \Gamma(f)d\mu = 0$ for the invariant measure μ of $\mathcal{L}$.

Definition 6. ***(Curvature-Dimension condition, (Ambrosio et al., 2015))*** *We can say that the infinitesimal generator $\mathcal{L}$ induces the $CD(\rho, \infty)$ curvature-dimension condition if it satisfies $\Gamma_1(f) \leq \rho\Gamma_2(f)$ for all $f \in C_0^\infty$.*

Because our diffusion operator generates a semi-group with respect to the Gibbs measure, the curvature-dimension condition can be calculated explicitly. Through simple calculations, the first-order ($c = 1$) diffusion carré du champ can be induced as follows:

$$\Gamma_1(f) = \left([\nabla f]^T \Sigma_\xi \nabla f\right)^2. \tag{13}$$

Similarly, the second-order ($c = 2$) diffusion carré du champ is calculated as follows:

$$\begin{aligned}\Gamma_2(f) &= \frac{1}{2}\left[\mathcal{L}\left(\Gamma_1(f^2)\right) - 2\Gamma_1\left(f, \mathcal{L}(f)\right)\right] \\ &= \mathbf{Tr}\left(\left[\Sigma_\xi \nabla^2 f\right]^2\right) + \left([\nabla f]^T \Sigma_\xi \nabla f\right)^2 = \mathbf{Tr}\left(\left[\Sigma_\xi \nabla^2 f\right]^2\right) + \Gamma_1(f),\end{aligned} \tag{14}$$

for an arbitrary $f \in C_0^\infty(\mathbb{R}^d)$. While $\mathbf{Tr}\left(\left[\Sigma \nabla^2 f\right]^2\right)$ is non-negative, we can infer that $\Gamma_1 \leq \Gamma_2$. In this case, the diffusion operator $\mathcal{L}$ defined in Lemma 1 induces the $CD(\rho = 1, \infty)$ curvature-dimension condition. For the other diffusion operators, please refer to (Bolley & Gentil, 2010).

Proposition 2. ***(Decay of Fisher information along a Markov semigroup, (Bakry et al., 2013).)*** *If we assume the curvature-dimension condition $CD(\rho, \infty)$, then $I(\mu_t|\mathcal{N}_\xi) \leq e^{-2\rho t} I(\mu|\mathcal{N}_\xi)$.*

The exponential decay of the Fisher information in Proposition 2 is a core property of the exponential decay of the Wasserstein distance.

5.3. Fokker-Plank equation, SDE

Definition 7. *(Over-damped Langevin Dynamics) We have*

$$dX_t = -\nabla\phi(X_t; \mathbf{m}_\xi)dt + \sqrt{2\tau^{-1}\Sigma_\xi} dW_t, \tag{15}$$

where $\phi\left(X_t; \mathbf{m}_\xi\right) = \frac{\tau}{2} d^2\left(X_t, \mathbf{m}_\xi\right)$, W_t denotes Brownian motion, and d denotes Euclidean distance. The particle X_t is distributed in $X_t \sim p_t$. The probability density $\lim_{t\to\infty} p(x, t)$ with respect to X_∞ converges to the Gaussian density $X_\infty = \sqrt{\Sigma_\xi}(Z + \mathbf{m}_\xi) \sim p_\infty(x) = q(x) \propto e^{-d(x, \mathbf{m}_\xi)^T \Sigma_\xi^{-1} d(x, \mathbf{m}_\xi)}$.

In classical SDE literature, it is stated that $\mathbb{E}\left[\sup_{0\leq t\leq T}\left|\hat{X}_t - X_t\right|\right] \leq G(N\varrho)^{-\frac{1}{2}}$, where $G(T)$ is some constant that depends only on T and $\hat{X}$ denotes the true solution of the SDE in (15). While the number of uncertain samples is greater than $N\varrho > 40$, our method exhibits acceptable convergence.

5.4. Gaussian Wasserstein Subspaces

It is known that the space of non-degenerate Gaussian measures (*i.e.*, covariance matrices are positive-definite) forms a subspace in the 2-Wasserstein space denoted as $\mathcal{W}_{2,g} \cong \mathbf{Sym}_d^+ \times \mathbb{R}^d$. Because the 2-Wasserstein space can be considered as a Riemannian manifold equipped with Riemannian metrics (Villani, 2008), $\mathcal{W}_{2,g}$ can be endowed with a Riemannian structure that also induces the Wasserstein metric ((McCann, 1997)). In the Riemannian sub-manifold of Gaussian measures, the geodesic between two points $\gamma(0) = \mathcal{N}_A$ and $\gamma(1) = \mathcal{N}_B$ is defined as follows (Malagò et al., 2018):

$$\gamma(\alpha) = \mathcal{N}_t = \mathcal{N}(\mathbf{m}(\alpha), \Sigma(\alpha)), \tag{16}$$

where $\mathbf{m}(\alpha) = (1-\alpha)\mathbf{m}_A + \alpha\mathbf{m}_B$ and $\Sigma(\alpha) = \left[(1-\alpha)\mathbf{I} + \alpha\mathcal{T}\right]\Sigma_A\left[(1-\alpha)\mathbf{I} + \alpha\mathcal{T}\right]$, where $\mathcal{T}\Sigma_A\mathcal{T} = \Sigma_B$. In Section 7 of the main paper, we set $(\mathbf{m}_A, \Sigma_A) \to (\mathbf{m}_\xi, \Sigma_\xi)$ and $(\mathbf{m}_B, \Sigma_B) \to (\mathbf{m}_{\xi_k}, \Sigma_{\xi_k})$. Regardless of how ξ is updated, the statistical information regarding the current certain measure ξ_k is considered in the detour Gaussian measure, which yields a much smoother geometric constraint on μ.

6. Proofs

Proposition. *The distributional normalization $\mathcal{F}$ maps μ into the certified robust region with controllable radius $\varepsilon_2 = K_2(\mu)e^{-t}$ (i.e., $\mathbb{B}_{\mathcal{N}_\xi}\left(K_2 e^{-t}\left(\mu\right)\right)$), where $K_2(\mu) > 0$ is a constant that depends on μ.*

Proof. We assume that the probability measure μ_t is absolutely continuous with respect to the detour Gaussian measure $\mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi) = \mathcal{N}_\xi$, $\mu_t \ll \mathcal{N}_\xi$. In this case, according to the Radon-Nikodym theorem, there is a corresponding unique probability density $q(t, x) = q_t(x) \in C_0^\infty$ such that $d\mu_t = q_t d\mathcal{N}_\xi$.

Lemma 2. *(**WI-inequality**, (Otto & Villani, 2000)) If the stationary state of μ_t with respect to P_t satisfies $\lim_{t\to\infty}\mathbb{E}_\mu[P_t f]=0$ for any $f\in C_0^\infty$, then the following inequality holds:*

$$\frac{d}{dt_+}\mathcal{W}_2(\mu,\mu_t)\leq\sqrt{I(\mu_t|\mathcal{N}_\xi)}. \tag{17}$$

By integrating both sides of the inequality in Lemma 2 with respect to $t\in(0,\infty)$, the following inequality can be obtained:

$$\mathcal{W}_2(\mu_t,\mathcal{N}_\xi)=\int_0^\infty\frac{d}{dt_+}\mathcal{W}_2(\mu_t,\mathcal{N}_\xi)dt\leq\int_0^\infty\sqrt{I(\mu_t|\mathcal{N}_\xi)}dt. \tag{18}$$

In the aforementioned inequality, we replace the Fisher information with the diffusion generator $\mathcal{L}$ as follows:

$$\begin{aligned}\mathcal{W}_2(\mu,\mathcal{N}_\xi)&\leq\int_0^\infty\sqrt{I(\mu_t|\mathcal{N}_\xi)}dt\\&=\int_0^\infty\sqrt{\int[P_tq]^{-1}\Gamma(P_tq)d\mathcal{N}_\xi}dt=\int_0^\infty\sqrt{\int\mathcal{L}(-\log P_tq)d\mu_t}dt.\end{aligned} \tag{19}$$

The second equality above is derived by leveraging the properties of the bilinear operator Γ ((Bakry et al., 2013; Villani, 2008)) with respect to the diffusion operator $\mathcal{L}$, which is defined as follows:

$$\int[P_tq]^{-1}\Gamma(P_tq)d\mathcal{N}_\xi=-\int\mathcal{L}(\log P_tq)q_td\mathcal{N}_\xi=\int\mathcal{L}(-\log P_tq)d\mu_t\geq 0. \tag{20}$$

For simplicity, we denote $|g|=g^+$ for any $g\in C_0^\infty$. According to Proposition 2, we can relate $\mathcal{F}_t\mu=\mu_t$ to its initial term $\mu=\mu_{t=0}$ as follows:

$$\begin{aligned}\int_0^\infty\sqrt{\int\mathcal{L}(-\log P_tq)(X)d[\mathcal{F}_t\mu](X)}dt&\leq\int_0^\infty\sqrt{e^{-2\rho t}\int\mathcal{L}\left(-\log P_{t=0}q\right)(X)d\mu(X)}dt\\&\leq\int_0^\infty\sqrt{e^{-2\rho t}\sup_{g\in C_0^\infty}\int\mathcal{L}^+g(Z)qd\mathcal{N}_\xi(Z)}dt\\&=\int_0^\infty\sqrt{e^{-2\rho t}}dt\sqrt{\sup_{g\in C_0^\infty}\int\mathcal{L}^+g(X)d\mu(X)}\\&=\rho^{-1}K_2(\mu).\end{aligned} \tag{21}$$

The second inequality is naturally induced, because the proposed objective function is defined to select the maximum elements over the set of functions $g\in C_0^\infty$ and $\mathcal{L}g\leq\mathcal{L}^+g$. If the integral interval is set to $(0,s)$, then we can induce $\mathcal{W}_2(\mu,\mathcal{F}_t\mu)\leq\frac{1}{\rho}(1-e^{-s})K_2(\mu)$. Our diffusion-operator induces $\rho=1$, which completes the proof. □

Proposition. *There is a scalar $0<\beta<\infty$ dependent on ξ such that the following inequality holds:*

$$\mathcal{W}_2(\xi,\mathcal{F}_t\mu)\leq\left[\sqrt{d\beta\lambda_{max}(\Sigma_\xi)}+\|\mathbb{E}_\xi Y\|_2\right]\vee\left[e^{-t}K_2(\mu)+K_2(\xi)\right]. \tag{22}$$

As a motivation for setting a detour measure to $\mathcal{N}_\xi$, we mentioned the natural property of the non-collapsing Wasserstein distance of $\mathcal{W}_2(\xi,\mathcal{N}_\xi)\neq 0$. However, it is unclear from a geometric perspective exactly how the upper bound (*i.e.*, $\mathcal{W}_2(\xi,\mathcal{N}_\xi)\leq ?$) can be induced based on the intrinsic statistics term (*i.e.*, ε_1). Specifically, in the situation where the covariance matrices of ξ and $\mathcal{N}_\xi$ are identical, it is difficult to determine a theoretical upper bound without additional tools. The first part of this proof focuses on resolving this important issue. The second part of the proof is naturally induced by Proposition 1. Please note that in the following proposition, parameter for Wasserstein moving average is set to $\alpha=0$ for clarity.

Proof. Before proceeding with the first part of the proof, we define a constant β as follows:

$$\beta = \sup_{1\leq j\leq d} \int_0^1 \frac{1}{s}\mathbb{E}_{Y_s} v_{s,j}^2(Y_s) ds. \tag{23}$$

If we assume a mild condition such that $\min_{s,j} \inf_{1\leq j\leq d} O(v_{s,j}) \geq O(\sqrt{s})$, then the integral term in β is finite and well-defined. This value will directly yield the upper bound of the Kullback–Leibler (KL) divergence of ξ. First, we introduce the following inequality.

Lemma 3. ***(de Bruijn's identity**, (Johnson & Suhov, 2001; Nourdin et al., 2014)) We let $Y \sim \xi$, $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote a standard Gaussian random variable, and let define $Y_s = \sqrt{s}Y + \sqrt{1-s}\Sigma_\xi^{\frac{1}{2}} Z$ with the score function defined as $v_s(x) = \nabla \log p_s(x)$ with respect to the random variable Y_s. Then, the following equality holds:*

$$\mathbf{KL}(\xi|\mathcal{N}(\mathbf{0}, \Sigma_\xi)) = \int_0^1 \mathbf{Tr}\left(\frac{1}{2s}\Sigma_\xi \mathbb{E}_{p_s \sim Y_s}[v_s(Y_s)v_s(Y_s)^T]\right) ds. \tag{24}$$

From (24), we can derive the relations between KL-divergence and the constant β defined earlier.

$$\begin{aligned} &\int_0^1 \frac{1}{2s}\mathbf{Tr}\left(\Sigma_\xi \mathbb{E}_x[v_s(Y_s)v_s(Y_s)^T])\right) ds \leq \int_0^1 \frac{1}{2s}\mathbf{Tr}\left(\Sigma_\xi \mathbb{E}_x[v_{s,i}v_{s,j}]_{i,j}^d)\right) ds \\ &\leq \int_0^1 \frac{1}{2}\lambda_{max}(\Sigma_\xi) \sum_{j=1}^d \mathbb{E}\left[\frac{v_{s,j}^2(Y_s)}{s}\right] ds \leq \frac{1}{2}\lambda_{max} \int_0^1 \sum_{j=1}^d \beta ds = \frac{1}{2}\lambda_{max}(\Sigma_\xi) d\beta. \end{aligned} \tag{25}$$

The second inequality holds based on the following element property of symmetric positive-definite matrices:

$$\mathbf{Tr}(AB) \leq \|A\|_{op} \mathbf{Tr}(B) = \lambda_{max}(A)\mathbf{Tr}(B), \quad \forall A, B \in \mathbf{Sym}_d^+. \tag{26}$$

It should be noted that because the distribution of ξ is compactly supported (*i.e.*, $\mathbf{supp}(q)$ is compact), the maximum eigenvalue of the covariance Σ_ξ is finite. The other relations are induced by the aforementioned definition. Next, we relate the KL-divergence and 2-Wasserstein distance naturally.

Definition 8. ***(Talagrand inequality for Gaussian measures**, (Otto & Villani, 2000)) For any non-degenerate Gaussian measure $\mathcal{N}$ with a mean 0, the following inequality is satisfied:*

$$\mathcal{W}_2(\xi, \mathcal{N}) \leq \sqrt{2\mathbf{KL}(\xi|\mathcal{N})}, \quad \forall \xi \in \mathcal{P}_2(\mathbb{R}^d). \tag{27}$$

By combining Definition 8 and (25), we can derive the following expression:

$$\mathcal{W}_2(\xi, \mathcal{N}(0, \Sigma_\xi)) \leq \sqrt{2\mathbf{KL}(\xi|\mathcal{N}(0, \Sigma_\xi))} \leq \sqrt{d\beta\lambda_{max}(\Sigma_\xi)} < \infty. \tag{28}$$

According to the triangle inequality for the 2-Wasserstein distance, we obtain:

$$\mathcal{W}_2(\xi, \mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi)) \leq \mathcal{W}_2(\xi, \mathcal{N}(0, \Sigma_\xi)) + \mathcal{W}_2(\mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi), \mathcal{N}(0, \Sigma_\xi)). \tag{29}$$

We investigated that the geodesic distance between two Gaussian measures having the same covariance is equivalent to the Euclidean distance between two means. Therefore, we can obtain the following equality:

$$\begin{aligned} \mathcal{W}_2(\mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi), \mathcal{N}(0, \Sigma_\xi)) &= \mathcal{W}_2(\iota_\#^{\mathbf{m}_\xi}[\mathcal{N}(0, \Sigma_\xi)], \mathcal{N}(0, \Sigma_\xi)) \\ &= \|\mathbf{m}_\xi - 0\|_2 = \|\mathbb{E}_\xi Y\|_2, \end{aligned} \tag{30}$$

where $\iota^{\mathbf{a}}(X) = X + \mathbf{a}$ for any vector $\mathbf{a} \in \mathbf{supp}(q)$. Now, by adding the two inequalities defined earlier, we can obtain

$$\mathcal{W}_2(\xi, \mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi)) \leq \|\mathbb{E}_\xi Y\|_2 + \sqrt{d\beta\lambda_{max}(\Sigma_\xi)}, \tag{31}$$

where it is easily shown that the upper-bound is only dependent on the statistical structure of ξ. Specifically, the term $\|\mathbb{E}_\xi Y\|_2$ represents the center of mass for a density of ξ and $\sqrt{d\beta\lambda_{max}(\Sigma_\xi)}$ is related to the covariance structure of ξ.

By applying Proposition 6 to both $\mathcal{F}_t\mu$ and ξ, we can easily recover (7) of the main paper as follows:

$$\begin{aligned}\mathcal{W}_2(\xi, \mathcal{F}_t\mu) \leq \varepsilon &= \mathcal{W}_2(\xi, \mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi)) + \mathcal{W}_2(\mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi), \mathcal{F}_t\mu) \\ &\leq \left(\left[\left\| \mathbb{E}_\xi Y \right\|_2 + \sqrt{d\beta\lambda_{max}(\Sigma_\xi)} \right] \wedge K_2(\xi) \right) + e^{-t}K_2(\mu) \\ &\leq \left[\sqrt{d\beta\lambda_{max}(\Sigma_\xi)} + \left\| \mathbb{E}_\xi Y \right\|_2 \right] \vee \left[e^{-t}K_2(\mu) + K_2(\xi) \right].\end{aligned} \tag{32}$$

The second inequality is easily obtained as $(a \wedge b) + c \leq a \vee (b + c)$ for any $a, b, c \geq 0$, which completes the proof. □

Because our detour measure is Gaussian, we have the following inequality for any $h \in C_0^\infty(\mathbb{R}^d)$:

$$\hat{K}_2(\mu) = \lim_{s \to 0} \frac{1}{s} \mathbb{E}_{X,Z \sim \mathcal{N}_\mathbf{I}} \left[h\left(e^{-s}X + \sqrt{1 - e^{-2s}}(\Sigma_\xi^{\frac{1}{2}} Z + \mathbf{m}_\xi) \right) - h(X) \right] \leq K_2(\mu) \tag{33}$$

where this equality holds if h is selected to induce a supremum over the set C_0^∞. For approximation, we simply consider $h(X) = \|X\|_2$ as a test function. In this case, the following inequality naturally holds: $\hat{\varepsilon} = \hat{K}_2(\xi) + \hat{K}_2(\mathcal{F}\mu) \leq K_2(\xi) + K_2(\mathcal{F}\mu) \leq K_1(\xi) \vee (K_2(\xi) + K_2(\mathcal{F}\mu)) = \varepsilon$. Thus, $\hat{\varepsilon}$ can be considered as an approximation of the theoretical upper bound ε suggested in Proposition 1. Subsequently, we investigate the effects of Wasserstein normalization based on $\hat{K}_2(\mu)$ in (33).

We explicitly introduce the detailed assumptions on μ as follows:

$$\mathbb{E}_{\mathcal{F}_{s^\star}\mu}[f^2] - [\mathbb{E}_{\mathcal{F}_{s^\star}\mu}[f]]^2 \leq (1 + \eta)\mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\mathbf{A}\nabla f^T \nabla f], \quad f \in C_0^\infty(\mathbb{R}^d) \tag{34}$$

This inequality renders the probabilistic assumptions for relating curvature-dimension condition and Wasserstein distance. This property is core to induce the explicit concentration inequality.

Proposition. *Let assume that the uncertain measure μ satisfy the inequality above. Then, there exists $\delta > 0$ such that the following concentration inequality for an uncertain measure holds:*

$$\hat{\mu}\Big(|\sigma - \mathbb{E}_\xi[\sigma]| \geq \delta \Big) \leq 6e^{-\frac{\sqrt{2}\delta^{\frac{3}{2}}}{K_2(\mu)}}, \tag{35}$$

where σ denotes the soft-max function.

Proof. Before proceeding with the main proof, we first prove the existence of $s^\star$. The limit of the interval with respect to η converges to a singleton $\{\infty\}$ as $I = \lim_{\eta \to 0}[\frac{1}{\eta}, \infty)$. In this case, (34) is the same as the Poincaré inequality for a Gaussian measure $\mathcal{N}_\xi$, which can be written as

$$\begin{aligned}\lim_{\eta \to 0} \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[f^2] - [\mathbb{E}_{\mathcal{F}_{s^\star}\mu}[f]]^2 &\leq \lim_{\eta \to 0}(1 + \eta)\mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\mathbf{A}\nabla f^T \nabla f] \\ &= \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\Sigma_\xi \nabla f^T \nabla f].\end{aligned} \tag{36}$$

While the Poincaré inequality in (36) is uniquely defined, we can find at least one value $s^\star$ satisfying (34). Let $X(t, w) = X_t(w)$ denote the stochastic process with respect to $q_t(x)$ defined in the proof of Definition 2 of the main paper. Additionally, let $c = \mathbb{E}_\xi[\sigma] - \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\sigma]$. Then, we can obtain the following inequality:

$$\begin{aligned}c = \mathbb{E}_\xi[\sigma] - \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\sigma] &= \kappa \left(\mathbb{E}_\xi \left[\frac{\sigma}{\kappa} \right] - \mathbb{E}_{\mathcal{F}_{s^\star}\mu} \left[\frac{\sigma}{\kappa} \right] \right) \leq \kappa \sup_{g \in \mathbf{Lip}_1} (\mathbb{E}_\xi g - \mathbb{E}_{\mathcal{F}_{s^\star}\mu} g) \\ \leq \kappa \mathcal{W}_1(\mathcal{F}_{s^\star}\mu, \xi) \leq \kappa \mathcal{W}_2(\mathcal{F}_{s^\star}\mu, \xi) \leq \frac{\kappa K_2(\mu)}{1 + \eta}.\end{aligned} \tag{37}$$

The first inequality is induced by the assumption regarding the κ-Lipschitzness of the function σ and the second inequality is induced by the Kantorovich-Rubinstein theorem. The third inequality is natural because $\mathcal{W}_a(\cdot, \cdot) \leq \mathcal{W}_b(\cdot, \cdot)$ for any $1 \leq a \leq b < \infty$. because (34) is equivalent to the Poincaré inequality for the measure $\mathcal{F}_{s^\star}\mu$, it satisfies the Bakry-emery curvature-dimension condition $CD(1 + \eta, \infty)$. Thus, as shown in the proof of Proposition 1 (*i.e.*, (21)), the last inequality is induced. Additionally, based on the concentration inequality of $\mathcal{F}_{s^\star}\mu$ [Proposition 4.4.2 (Bakry et al., 2013)], we can derive the following probability inequality:

$$\mathcal{F}_{s^\star}\mu\left[\sigma(X_{s^\star}(w)) \geq \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\sigma] + \delta\right] \leq 3e^{-\frac{\delta}{\sqrt{1+\eta\kappa}}}, \tag{38}$$

where the Poincaré constant for $\mathcal{F}_{s^\star}\mu$ is naturally $1+\eta$ and $\|\sigma\|_{Lip} = \kappa$. Next, we will derive the desired form from (38). First, we introduce the following inequality.

$$\sigma(X_{s^\star}) \geq \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\sigma] + \delta \geq \mathbb{E}_{\xi}[\sigma] + \delta - \frac{\kappa}{1+\eta}K_2 \tag{39}$$

The last inequality is directly induced by (37) because $-c \geq -\frac{\kappa}{1+\eta}K_2$. While η, κ, and K_2 are constants with respect to w, the following set inclusion can be obtained naturally:

$$\mathcal{S}_1 = \{w : \sigma(X_{s^\star}(w)) \geq \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\sigma] + \delta\} \supseteq \{w : \sigma(X_{s^\star}(w)) \geq \mathbb{E}_{\xi}[\sigma] + \delta - \frac{\kappa}{1+\eta}K_2\} = \mathcal{S}_2. \tag{40}$$

For the modified version of the original probability inequality, we take probability measure $\mathcal{F}_{s^\star}\mu[\cdot]$ for the sets $\mathcal{S}_1, \mathcal{S}_2$, which is defined as

$$\begin{aligned} 3e^{-\frac{\delta}{\sqrt{1+\eta\kappa}}} &\geq \mathcal{F}_{s^\star}\mu\left(\{w : \sigma(X_{s^\star}(w)) \geq \mathbb{E}_{\mathcal{F}_{s^\star}\mu}[\sigma] + \delta\}\right) \\ &\geq \mathcal{F}_{s^\star}\mu\left(\{w : \sigma(X_{s^\star}(w)) \geq \mathbb{E}_{\xi}[\sigma] + \delta - \frac{\kappa}{1+\eta}K_2\}\right). \end{aligned} \tag{41}$$

The concentration inequality around $\mathbb{E}_{\xi}[\sigma]$ is obtained by combining the inequalities induced by σ and $-\sigma$ as follows:

$$\begin{aligned} &\frac{1}{2}\mathcal{F}_{s^\star}\mu\left(\bigcup_{h\in\{\sigma,-\sigma\}}\{w : h(X_{s^\star}(w)) - \mathbb{E}_{\xi}[h] \geq \pm\left(\delta - \frac{\kappa}{1+\eta}K_2\}\right)\right) \\ &= \mathcal{F}_{s^\star}\mu\left(\{w : |\sigma(X_{s^\star}(w)) - \mathbb{E}_{\xi}[\sigma]| \geq \delta - \frac{\kappa}{1+\eta}K_2\}\right) \leq 6e^{-\frac{\delta}{\sqrt{1+\eta\kappa}}}. \end{aligned} \tag{42}$$

The inequality in (42) is the general form containing the relation between the upper bound of the probability and (η, κ, K_2). While this form is quite complicated and highly technical, we choose not to present all the detailed expressions of (42) in the main paper. Rather than that, we re-write it in a much simplified form for clarity. Specifically, by setting $\kappa K_2/(1+\eta) = 0.5\delta$ and rescaling δ to 2δ, the aforementioned inequality in (42) can be converted into the following simpler form:

$$\mathcal{F}_{s^\star}\mu\left(\{w : |\sigma(X_{s^\star}(w)) - \mathbb{E}_{\xi}[l]| \geq \delta\right) \leq 6e^{-\frac{\sqrt{2}\delta^{\frac{3}{2}}}{\kappa K_2}}. \tag{43}$$

□

Finally, if we set $\sigma = \mathbf{Softmax}$, then the Lipschitz constant is induced as $\kappa = 1$. This proof is completed by setting $s^\star := T$.

References

Ambrosio, L., Gigli, N., Savaré, G., et al. Bakry–émery curvature-dimension condition and riemannian ricci curvature bounds. *The Annals of Probability*, 43(1):339–404, 2015.

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and Mcguinness, K. Unsupervised label noise modeling and loss correction. In *ICML*, 2019.

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Bolley, F. and Gentil, I. Phi-entropy inequalities for diffusion semigroups. *Journal de mathématiques pures et appliquées*, 93(5):449–473, 2010.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Johnson, O. and Suhov, Y. Entropy and random vectors. *Journal of Statistical Physics*, 104:145–165, 01 2001. doi: 10.1023/A:1010353526846.

Lee, K., Yun, S., Lee, K., Lee, H., Li, B., and Shin, J. Robust inference via generative classifiers for handling noisy labels. In *ICML*, 2019.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019.

Malagò, L., Montrucchio, L., and Pistone, G. Wasserstein riemannian geometry of gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Nourdin, I., Peccati, G., and Swan, Y. Entropy and the fourth moment phenomenon. *Journal of Functional Analysis*, 266(5): 3170–3207, 2014.

Otto, F. and Villani, C. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *ICML*, 2018.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.