A Appendix of Proofs

A.1 Proof of Thm.3.2

Theorem 3.2. *By choosing KL divergence $D_{KL}(Q||Q_0) = \int Q \log \frac{Q}{Q_0} dx$, optimizing CL-DRO (*cf. *Eqn. 3) is equivalent to optimizing CL (InfoNCE,*cf. *Eqn. 1):*

$$\begin{aligned} \mathcal{L}^{KL}_{\textit{CL-DRO}} &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1} \max_{Q\in\mathbb{Q}} \{\mathbb{E}_Q[f_\theta] - \alpha[D_{KL}(Q||Q_0) - \eta] + \eta_1(\mathbb{E}_{Q_0}[\frac{Q}{Q_0}] - 1)\} \\ &= -\mathbb{E}_{P_0}\big[\alpha^* \log \frac{e^{f_\theta/\alpha^*}}{\mathbb{E}_{Q_0}[e^{f_\theta/\alpha^*}]}\big] + Constant = \alpha^* \mathcal{L}_{\textit{InfoNCE}} + Constant \end{aligned} \tag{4}$$

where α, η_1 represent the Lagrange multipliers, and the optimal α^* finally serves as the temperature τ in CL.

Proof. To complete the proof, we start with giving some important notations and theorem.

Definition A.1 (ϕ-divergence [13])**.** For any convex funtion ϕ with $\phi(1) = 0$, the ϕ-divergence between Q and Q_0 is:

$$D_\phi(Q||Q_0) \coloneqq \mathbb{E}_{Q_0}[\phi(dQ/dQ_0)] \tag{13}$$

where $D_\phi(Q||Q_0) = \infty$ if P is not absolutely continuous with respect to Q_0. Specially, when $\phi(x) = x\log x - x + 1$, ϕ-divergence degenerates to the well-known KL divergence.

Definition A.2 (Convex conjugate [9])**.** We consider a pair (A, B) of topological vector spaces and a bilinear form $\langle\cdot,\cdot\rangle \to \mathbb{R}$ such that $(A, B, \langle\cdot,\cdot\rangle)$ form a dual pair. For a convex function $f : \mathbb{R} \to \mathbb{R}$, $dom f \coloneqq \{x \in \mathbb{R} : f(x) < \infty\}$ is the effective domain of f. The convex conjugate, also known as the Legendre-Fenchel transform, of $f : A \to \mathbb{R}$ is the function $f^* : B \to \mathbb{R}$ defined as

$$f^*(b) = \sup_a \{ab - f(a)\}, \quad b \in B \tag{14}$$

Theorem A.3 (Interchange of minimization and integration [2])**.** *Let $(\Omega, \mathcal{F})$ be a measurable space equipped with σ-algebra $\mathcal{F}$, $L^p(\Omega, \mathcal{F}, P)$ be the linear space of measurable real valued functions $f : \Omega \to \mathbb{R}$ with $||f||_p < \infty$, and let $\mathcal{X} \coloneqq L^p(\Omega, \mathcal{F}, P)$, $p \in [1, +\infty]$. Let $g : \mathbb{R} \times \Omega \to \mathbb{R}$ be a normal integrand, and define on $\mathcal{X}$. Then,*

$$\min_{x\in\mathcal{X}} \int_\Omega g(x(\omega), \omega)\, dP(\omega) = \int_\Omega \min_{s\in\mathbb{R}} g(s, \omega)\, dP(\omega) \tag{15}$$

To ease the derivation, we denote the likelihood raito $L(x, y) = Q(x, y)/Q_0(x, y)$. Note that the ϕ-divergence between Q and Q_0 is constrained, and thus $L(.)$ is fine definition. For brevity, we usually short $L(x, y)$ as L. And in terms of definition A.1 of ϕ-divergence, the expression of CL-DRO becomes:

$$\mathcal{L}^\phi_{\text{CL-DRO}} = -\mathbb{E}_{P_0}[f_\theta] + \max_L \mathbb{E}_{Q_0}[f_\theta L] \qquad \textit{s.t. } \mathbb{E}_{Q_0}[\phi(L)] \leq \eta \tag{16}$$

Note that $\mathbb{E}_{Q_0}[f_\theta L]$ and $\mathbb{E}_{Q_0}[\phi(L)]$ are both convex in L. We use the Lagrangian function solver:

$$\begin{aligned} \mathcal{L}^\phi_{\text{CL-DRO}} &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1} \max_L \{\mathbb{E}_{Q_0}[f_\theta L] - \alpha[\mathbb{E}_{Q_0}[\phi(L)] - \eta] + \eta_1(\mathbb{E}_{Q_0}[L] - 1)\} \\ &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1} \big\{\alpha\eta - \eta_1 + \alpha \max_L \{\mathbb{E}_{Q_0}[\frac{f_\theta + \eta_1}{\alpha} L - \phi(L)]\}\big\} \\ &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1} \big\{\alpha\eta - \eta_1 + \alpha \mathbb{E}_{Q_0}[\max_L \{\frac{f_\theta + \eta_1}{\alpha} L - \phi(L)\}]\big\} \\ &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1} \big\{\alpha\eta - \eta_1 + \alpha \mathbb{E}_{Q_0}[\phi^*(\frac{f_\theta + \eta_1}{\alpha})]\big\} \end{aligned} \tag{17}$$

The first equality holds due to the strong duality [3]. The second equality is a re-arrangement for optmizing L. The third equation follows by the Thm. A.3. The last equality is established based on the definition of convex conjugate A.2. When we choose KL-divergence, we have $\phi_{KL}(x) = x\log x - x + 1$. It can be deduced that $\phi^*_{KL}(x) = e^x - 1$. Then, we have:

$$\begin{aligned}
\mathcal{L}^{KL}_{\text{CL-DRO}} &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1}\left\{\alpha\eta - \eta_1 + \alpha\mathbb{E}_{Q_0}[\phi^*(\frac{f_\theta+\eta_1}{\alpha})]\right\} \\
&= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0,\eta_1}\left\{\alpha\eta - \eta_1 + \alpha\mathbb{E}_{Q_0}[e^{\frac{f_\theta+\eta_1}{\alpha}} - 1]\right\} \\
&= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0}\left\{\alpha\eta + \alpha\log\mathbb{E}_{Q_0}[e^{\frac{f_\theta}{\alpha}}]\right\} \\
&= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha\geq 0}\left\{\alpha\eta + \alpha\log\mathbb{E}_{Q_0}[e^{\frac{f_\theta}{\alpha}}]\right\} \\
&= -\mathbb{E}_{P_0}\left[\alpha^*\log\frac{e^{f_\theta/\alpha^*}}{\mathbb{E}_{Q_0}[e^{f_\theta/\alpha^*}]}\right] + \alpha\eta \\
&= \alpha^*\mathcal{L}_{\text{InfoNCE}} + Constant
\end{aligned} \tag{18}$$

Here the α^* represents the optimal value of $\min_{\alpha\geq 0}\left\{\alpha\eta + \alpha\log\mathbb{E}_{Q_0}[e^{\frac{f_\theta}{\alpha}}]\right\}$. $\square$

A.2 Proof of Thm.3.3

Theorem 3.3. *[Generalization Bound] Let $\widehat{\mathcal{L}}_{\textit{InfoNCE}}$ be an estimation of InfoNCE with N negative samples. Then if $Q^{\textit{ideal}}$ satisfied $D_{KL}(Q^{\textit{ideal}}||Q_0) \leq \eta$, we have that with probability at least $1-\rho$:*

$$\mathcal{L}_{\textit{unbiased}} \leq \tau\widehat{\mathcal{L}}_{\textit{InfoNCE}} + \mathcal{B}(\rho, N, \tau) \tag{5}$$

where $\mathcal{B}(\rho, N, \tau) = \frac{1}{N-1+\exp(\frac{1}{\tau})}\sqrt{\frac{N\exp(\frac{2}{\tau})\log(\frac{1}{\rho})}{2}}$.

Here we simply disregard the constant term present in Eqn. 4 as it does not impact optimization, and omit the error from the positive instances.

Proof. Before detailing the proof process, we first introduce a pertinent theorem:

Theorem A.4 (McDiarmid's inequality [10])**.** *Let $X_1, \cdots, X_n$ be independent random variables, where X_i has range $\mathcal{X}$. Let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_n)$-bounded difference property: for every $i = 1, \ldots, n$ and every $(x_1, \ldots, x_n), (x'_1, \ldots, x'_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ that differ only in the i-th coordinate ($x_j = x'_j$ for all $j \neq i$), we have $|f(x_1, \ldots, x_n) - f(x'_1, \ldots, x'_n)| \leq c_i$. For any $\epsilon > 0$,*

$$\mathbb{P}(f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)] \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^N c_i^2}) \tag{19}$$

Now we delve into the proof. As Q^{ideal} satisfies $D_{KL}(Q||Q_0) \leq \eta$, we can bound $\mathcal{L}_{unbiased}$ with:

$$\begin{aligned}
\mathcal{L}_{unbiased} &= -\mathbb{E}_{P_0}[f_\theta] + \mathbb{E}_{Q^{ideal}}[f_\theta] \\
&\leq -\mathbb{E}_{P_0}[f_\theta] + \max_{D_{KL}(Q|Q_0)\leq\eta}\mathbb{E}_Q[f_\theta] \\
&= \mathcal{L}^{KL}_{\text{CL-DRO}}
\end{aligned} \tag{20}$$

where Q^{ideal}, Q^* denotes the ideal negative distribution and the worst-case distribution in CL-DRO. From the Thm.3.2, we have the equivalence between InfoNCE and CL-DRO. Thus here we choose CL-DRO for analyses. Suppose we have N negative samples, and for any pair of samples $(x_i, y_i), (x_j, y_j)$, we have the following bound:

$$|Q^*(x_i, y_i)f_\theta(x_i, y_i) - Q^*(x_j, y_j)f_\theta(x_j, y_j)| \leq \sup_{(x,y)\sim Q_0}|Q^*(x,y)f_\theta(x,y)| \leq \frac{\exp\left(\frac{1}{\tau}\right)}{N-1+\exp\left(\frac{1}{\tau}\right)} \tag{21}$$

where the first inequality holds as $Q^*(x,y)f_\theta(x,y) > 0$. The second inequality holds based on the expression of $Q^* = Q_0\frac{\exp[f_\theta/\tau]}{E_{Q_0}\exp[f_\theta/\tau]}$ (refer to Appendix A.6). Suppose $f_\theta \in [M_1, M_2]$, the upper bound of $\sup_{(x,y)\sim Q_0} |Q^*(x,y)f_\theta(x,y)|$ arrives if $f_\theta(x,y) = M_2$ for the sample (x,y) and $f_\theta(x,y) = M_1$ for others. We have $\sup_{(x,y)\sim Q_0} |Q^*(x,y)f_\theta(x,y)| \leq \frac{M_2 \exp((M_2-M_1)/\tau)}{N-1+\exp((M_2-M_1)/\tau)}$. In this work, for brevity, here we simply consider $M_1 = 0, M_2 = 1$ for analyses. It shares the common properties with the general interval $[M_1, M_2]$.

By using McDiarmid's inequality in Thm A.4,for any ϵ, we have:

$$\begin{aligned} &\mathbb{P}[(\mathcal{L}^{KL}_{\text{CL-DRO}} - \tau\widehat{\mathcal{L}}_{\text{InfoNCE}}) \geq \epsilon] \\ \leq &\exp\big(\frac{-2\epsilon^2(N-1+\exp(\frac{1}{\tau}))^2}{N\exp(\frac{2}{\tau})}\big) \end{aligned} \tag{22}$$

Let

$$\rho = \exp(\frac{-2\epsilon^2(N-1+\exp(\frac{1}{\tau}))^2}{N\exp(\frac{2}{\tau})}) \tag{23}$$

we get:

$$\epsilon = \frac{1}{N-1+\exp(\frac{1}{\tau})}\sqrt{\frac{N\exp(\frac{2}{\tau})\log(\frac{1}{\rho})}{2}} \tag{24}$$

Thus, for $\forall\rho \in (0,1)$, we conclude that with probability at least $1-\rho$.

$$\mathcal{L}_{\text{unbiased}} \leq \widehat{\mathcal{L}}_{\text{InfoNCE}} + \frac{1}{N-1+\exp(\frac{1}{\tau})}\sqrt{\frac{N\exp(\frac{2}{\tau})\log(\frac{1}{\rho})}{2}} \tag{25}$$

□

A.3 Proof of Coro.3.4

Corollary 3.4. *[The optimal α - Lemma 5 of [6]] The value of the optimal α (*i.e.*, τ) can be approximated as follow:*

$$\tau \approx \sqrt{\mathbb{V}_{Q_0}[f_\theta]/2\eta}. \tag{6}$$

where $\mathbb{V}_{Q_0}[f_\theta]$ denotes the variance of f_θ under the distribution Q_0.

Proof. While Corollary 3.4 has already been proven in [6], we present a brief outline of the proof here for the sake of completeness and to ensure that our article is self-contained. To verify the relationship between τ and η, we could utilize the approximate expression of InfoNCE (*cf.* Eqn. 29) and focus on the first order conditions for τ. In detail, we have:

$$-\mathbb{E}_{P_0}[f_\theta] + \inf_{\alpha\geq 0}\{\mathbb{E}_{Q_0}[f_\theta] - \frac{1}{2\alpha}\frac{1}{\phi^{(2)}(1)}\mathbb{V}_{Q_0}[f_\theta] - \alpha\eta\}$$

To find the optimal value of α (or equivalently, τ), we differentiate the above equation and set it to 0. This yields a fixed-point equation

$$\tau = \sqrt{\frac{\mathbb{V}_{Q_0}[f_\theta]}{2\eta}}$$

The corollary gets proved. □

A.4 Proof of Thm.3.5

Theorem 3.5. *Given any ϕ-divergence, the corresponding CL-DRO objective could be approximated as a mean-variance objective:*

$$\mathcal{L}^{\phi}_{\textit{CL-DRO}}(f_\theta) \approx -\mathbb{E}_{P_0}[f_\theta] + (\mathbb{E}_{Q_0}[f_\theta] + \frac{1}{2\tau}\frac{1}{\phi^{(2)}(1)}\cdot\mathbb{V}_{Q_0}[f_\theta]) \tag{7}$$

where $\phi^{(2)}(1)$ denotes the the second derivative value of $\phi(\cdot)$ at point 1, and $\mathbb{V}_{Q_0}[f_\theta]$ denotes the variance of f under the distribution Q_0.

Specially, if we consider KL divergence, the approximation transforms:

$$\mathcal{L}^{KL}_{\textit{CL-DRO}}(f_\theta) \approx -\mathbb{E}_{P_0}[f_\theta] + (\mathbb{E}_{Q_0}[f_\theta] + \frac{1}{2\tau}\mathbb{V}_{Q_0}[f_\theta]) \tag{8}$$

Proof. We start with introducing a useful lemma.

Lemma A.5 (Lemma A.2 of [8]). *Suppose that $\phi : \mathbb{R} \to \mathbb{R} \bigcup \{+\infty\}$ is a closed, convex function such that $\phi(z) \geq \phi(1) = 0$ for all z, is two times continuously differentiable around $z = 1$, and $\phi(1) > 0$, Then*

$$\begin{aligned} \phi^*(\zeta) &= \max_z \{z\zeta - \phi(z)\} \\ &= \zeta + \frac{1}{2!}[\frac{1}{\phi''(1)}]\zeta^2 + o(\zeta^2) \end{aligned} \tag{26}$$

Note that most of the ϕ-divergences [13] (*e.g.,* KL divergence, Cressie-Read divergence, Burg entropy, J-divergence, χ^2-distance, modified χ^2-distance, and Hellinger distance) satisfy the smoothness conditions. When $n = 2$, $\phi^*[\zeta] \approx \zeta + \frac{1}{2}[\frac{1}{\phi^{(2)}(1)}]\zeta^2$. Substituting this back to Eqn.17 we have:

$$\begin{aligned} \mathcal{L}^{\phi}_{\text{CL-DRO}} &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha \geq 0, \eta_1} \{\alpha\eta - \eta_1 + \alpha\mathbb{E}_{Q_0}[\phi^*(\frac{f_\theta + \eta_1}{\alpha})]\} \\ &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha \geq 0, \eta_1} \{\alpha\eta - \eta_1 + \alpha\mathbb{E}_{Q_0}[\frac{f_\theta + \eta_1}{\alpha} + \frac{1}{2}\frac{1}{\phi^{(2)}(1)}(\frac{f_\theta + \eta_1}{\alpha})^2]\} \\ &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha \geq 0, \eta_1} \{\alpha\eta - \eta_1 + \mathbb{E}_{Q_0}[f_\theta + \eta_1 + \frac{1}{2}\frac{1}{\phi^{(2)}(1)\alpha}(f_\theta + \eta_1)^2]\} \\ &= -\mathbb{E}_{P_0}[f_\theta] + \min_{\alpha \geq 0, \eta_1} \{\alpha\eta + \mathbb{E}_{Q_0}[f_\theta + \frac{1}{2}\frac{1}{\phi^{(2)}(1)\alpha}(f_\theta + \eta_1)^2]\} \end{aligned} \tag{27}$$

If we differentiate it *w.r.t.* η_1:

$$\frac{\partial\{\alpha\eta + \mathbb{E}_{Q_0}[f_\theta + \frac{1}{2}\frac{1}{\phi^{(2)}(1)\alpha}(f_\theta + \eta_1)^2]\}}{\partial\eta_1} = 0 \tag{28}$$

we have $\eta_1^* = -\mathbb{E}_{Q_0}[f_\theta]$, and the objective transforms into:

$$\begin{aligned} &-\mathbb{E}_{P_0}[f_\theta] + \inf_{\alpha \geq 0, \eta_1} \{\alpha\eta + \mathbb{E}_{Q_0}[f_\theta + \frac{1}{2}\frac{1}{\phi^{(2)}(1)\alpha}(f_\theta + \eta_1)^2]\} \\ =& -\mathbb{E}_{P_0}[f_\theta] + \inf_{\alpha \geq 0}\{\mathbb{E}_{Q_0}[f_\theta + \frac{1}{2\alpha}\frac{1}{\phi^{(2)}(1)}(f_\theta - \mathbb{E}_{Q_0}[f_\theta])^2] - \alpha\eta\} \\ =& -\mathbb{E}_{P_0}[f_\theta] + \inf_{\alpha \geq 0}\{\mathbb{E}_{Q_0}[f_\theta] + \frac{1}{2\alpha}\frac{1}{\phi^{(2)}(1)}\mathbb{V}_{Q_0}[f_\theta] - \alpha\eta\} \end{aligned} \tag{29}$$

Choosing KL-divergence, we have $\phi^{(2)}(1) = 1$. Substituting $\alpha^*(\tau)$ into Eqn. 29 and ignoring the constant $\alpha\eta$:

$$-\mathbb{E}_{P_0}[f_\theta] + \mathbb{E}_{Q_0}[f_\theta] + \frac{1}{2\tau}\mathbb{V}_{Q_0}[f_\theta]$$

Then Theorem gets proved. □

A.5 Proof of Thm.4.2

Theorem 4.2. *For distributions P, Q such that $P \ll Q$, let $\mathcal{F}$ be a set of bounded measurable functions. Let CL-DRO draw positive and negative instances from P and Q, marked as $\mathcal{L}^{\phi}_{\textit{CL-DRO}}(P, Q)$. Then the CL-DRO objective is the tight variational estimation of ϕ-divergence. In fact, we have:*

$$D_\phi(P||Q) = \sup_{f \in \mathcal{F}} -\mathcal{L}^{\phi}_{\textit{CL-DRO}}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \min_{\lambda \in \mathbb{R}}\{\lambda + \mathbb{E}_Q[\phi^*(f - \lambda)]\} \tag{10}$$

Here, the choice of ϕ in CL-DRO corresponds to the probability measures in $D_\phi(P||Q)$.

Proof. Regarding this theorem, our proof primarily relies on the variational representation of ϕ-divergence and optimized certainty equivalent (OCE) risk. Towards this end, we start to introduce the basic concepts:

Definition A.6 (OCE [2]). Let X be a random variable and let u be a convex, lower-semicontinuous function satisfies $u(0) = 0, u^*(1) = 0$, then optimized certainty equivalent (OCE) risk $\rho(X)$ is defines as:

$$\rho(X) = \inf_{\lambda \in \mathbb{R}} \{\lambda + \mathbb{E}[u(f - \lambda)]\} \tag{30}$$

OCE is a type of risk measure that is widely used by both practitioners and academics [1, 2]. With duality theory, its various properties have been inspiring in our study of DRO.

Definition A.7 (Variational formulation).

$$D_\phi(P||Q) \coloneqq \sup_{f \in \mathcal{F}} \{\mathbb{E}_P[f] - \mathbb{E}_Q[\phi^*(f)]\} \tag{31}$$

where the supremum is taken over all bounded real-valued measurable functions $\mathcal{F}$ defined on $\mathcal{X}$.

Note that in order to keep consistent with the definition of CL-DRO, we transform Eqn.10 to :

$$D_\phi(P||Q) = \sup_{f \in \mathcal{F}} -\mathcal{L}^\phi_{\text{CL-DRO}}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \inf_{\eta_1 \in \mathbb{R}} \{-\eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\} \tag{32}$$

Our proof for this theorem primarily relies on utilizing OCE risk as a bridge and can be divided into two distinct steps:

Step 1: $\sup_{f \in \mathcal{F}} -\mathcal{L}^\phi_{\text{CL-DRO}}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \inf_{\eta_1 \in \mathbb{R}} \{-\eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\}$.

Step 2: $D_\phi(P||Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \inf_{\eta_1 \in \mathbb{R}} \{-\eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\}$

1. **We show that** $\sup_{f \in \mathcal{F}} -\mathcal{L}^\phi_{\textbf{CL-DRO}}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \inf_{\eta_1 \in \mathbb{R}} \{-\eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\}$**.**

$$\begin{aligned}
-\mathcal{L}^\phi_{\text{CL-DRO}} &= \mathbb{E}_P[f] - \inf_{\alpha \geq 0, \eta_1} \sup_L \{\mathbb{E}_Q[fL] - \alpha[\mathbb{E}_Q[\phi(L)] - \eta] + \eta_1(\mathbb{E}_Q[L] - 1)\} \\
&= \mathbb{E}_P[f] - \inf_{\alpha \geq 0, \eta_1} \big\{\alpha\eta - \eta_1 + \alpha\mathbb{E}_Q[\phi^*(\frac{f + \eta_1}{\alpha})]\big\} \\
&= \mathbb{E}_P[f] - \inf_{\eta_1} \big\{\alpha^*\eta - \eta_1 + \alpha^*\mathbb{E}_Q[\phi^*(\frac{f + \eta_1}{\alpha^*})]\big\} \\
&= \mathbb{E}_P[f] - \inf_{\eta_1} \big\{ - \eta_1 + \alpha^*\mathbb{E}_Q[\phi^*(\frac{f + \eta_1}{\alpha^*})] + Constant\big\}
\end{aligned} \tag{33}$$

When $\alpha^* = 1$, step 1 gets proved.

2. **We show that** $D_\phi(P||Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \inf_{\eta_1 \in \mathbb{R}} \{-\eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\}$**.**

Firstly, we transform $\mathbb{E}_P[f]$ to $\mathbb{E}_Q[f\frac{dP}{dQ}]$ as:

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \inf_{\eta_1} \big\{ - \eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\big\} \\
=&\sup_{f \in \mathcal{F}} \mathbb{E}_Q[f\frac{dP}{dQ}] - \inf_{\eta_1} \big\{ - \eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\big\}
\end{aligned} \tag{34}$$

Let $f + \eta_1 = Y$, we have:

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \mathbb{E}_Q[f\frac{dP}{dQ}] - \inf_{\eta_1} \big\{ - \eta_1 + \mathbb{E}_Q[\phi^*(f + \eta_1)]\big\} \\
=&\sup_{Y \in \mathcal{F}} \inf_{\eta_1} \mathbb{E}_Q[(Y - \eta_1)\frac{dP}{dQ}] - \big\{ - \eta_1 + \mathbb{E}_Q[\phi^*(Y)]\big\} \\
=&\sup_{Y \in \mathcal{F}} \mathbb{E}_Q[Y\frac{dP}{dQ} - \phi^*(Y)] + \inf_{\eta_1} \eta_1\mathbb{E}_Q[1 - \frac{dP}{dQ}] \\
=&\sup_{Y \in \mathcal{F}} \mathbb{E}_Q[Y\frac{dP}{dQ} - \phi^*(Y)] + 0
\end{aligned} \tag{35}$$

The first equality follows from replacing $f+\theta_1$ with Y. The second equality is a re-arrangement for optmizing η_1. The third equation holds as $\mathbb{E}_Q[1-\frac{dP}{dQ}]=0$.

Applying Thm. A.3, the last supremum reduces to:

$$\begin{aligned}
&\sup_{Y\in\mathcal{F}}\mathbb{E}_Q[Y\frac{dP}{dQ}-\phi^*(Y)] \\
=&\mathbb{E}_Q[\sup_{Y\in\mathcal{F}}\{Y\frac{dP}{dQ}-\phi^*(Y)\}] \\
=&\mathbb{E}_Q[\phi^{**}(\frac{dP}{dQ})] \\
=&\mathbb{E}_Q[\phi(\frac{dP}{dQ})] \\
=&D_\phi(P||Q)
\end{aligned} \tag{36}$$

where the last equality follows from the fact that $\phi^{**}=\phi$. This concludes the proof.

□

A.6 Proof of Q^*

Proof. From theorm3.2, CL-DRO can be rewrriten as:

$$\begin{aligned}
\mathcal{L}^{\phi}_{\text{CL-DRO}} &= -\mathbb{E}_{P_0}[f_\theta]+\min_{\eta_1}\left\{\alpha^*\eta-\eta_1+\alpha^*\mathbb{E}_{Q_0}[\max_L\{\frac{f_\theta+\eta_1}{\alpha^*}L-\phi(L)\}]\right\} \\
&= -\mathbb{E}_{P_0}[f_\theta]+\min_{\eta_1}\left\{\alpha^*\eta-\eta_1+\alpha^*\mathbb{E}_{Q_0}[\phi^*(\frac{f_\theta+\eta_1}{\alpha^*})]\right\}
\end{aligned} \tag{37}$$

For the inner optimzation, we can draw the optimal L for $\max_L\{\frac{f_\theta+\eta_1}{\alpha^*}L-\phi(L)\}$ as:

$$L=e^{\frac{f_\theta+\eta_1}{\alpha^*}} \tag{38}$$

For the outer optimization, we can draw the optimal η_1 for $\min_{\eta_1}\left\{\alpha^*\eta-\eta_1+\alpha^*\mathbb{E}_{Q_0}[e^{\frac{f_\theta+\eta_1}{\alpha^*}}-1]\right\}$ as

$$\eta_1=-\alpha^*\log\mathbb{E}_{Q_0}e^{\frac{f_\theta}{\alpha^*}} \tag{39}$$

Then we plug Eqn. 39 into Eqn. 38.

$$L=\frac{e^{\frac{f_\theta}{\alpha^*}}}{\mathbb{E}_{Q_0}[e^{\frac{f_\theta}{\alpha^*}}]} \tag{40}$$

Based on the definition of L, we can derive the expression for Q^*:

$$Q^*=\frac{e^{\frac{f_\theta}{\alpha^*}}}{\mathbb{E}_{Q_0}[e^{\frac{f_\theta}{\alpha^*}}]}Q_0 \tag{41}$$

□

B Experiments

Figure 5 shows PyTorch-style pseudocode for the standard objective, the adjusted InfoNCE objective. The proposed adjusted reweighting loss is very simple to implement, requiring only two extra lines of code compared to the standard objective.

```
# pos      : exp of inner products for positive examples
# neg      : exp of inner products for negative examples
# N        : number of negative examples
# t        : temperature scaling
# mu       : center position
# sigma    : height scale

#InfoNCE
standard_loss = -log(pos.sum() / (pos.sum() + neg.sum()))

#ADNCE
weight=1/(sigma * sqrt(2*pi)) * exp( -0.5 * ((neg-mu)/sigma)**2 )
weight=weight/weight.mean()
Adjusted_loss = -log(pos.sum() / (pos.sum() + (neg * weight.detach() ).sum())
    )
```

Figure 5: Pseudocode for our proposed adjusted InfoNCE objective, as well as the original NCE contrastive objective. The implementation of our adjusted reweighting method only requires two additional lines of code compared to the standard objective.

Table 6: hyperparameters setting on each datasets.

DATASETS	CIFAR10	STL10	CIFAR100
BEST τ	{ 0.1, 0.2, 0.3, **0.4**, 0.5, 0.6 }	{ 0.1, **0.2**, 0.3, 0.4, 0.5, 0.6 }	{ 0.1, 0.2, **0.3**, 0.4, 0.5, 0.6 }
μ	{ 0.5, 0.6, **0.7**, 0.8, 0.9 }	{ 0.5, 0.6, 0.7, **0.8**, 0.9 }	**0.5**, 0.6, 0.7, 0.8, 0.9 }
σ	{**0.5**, 1.0 }	{0.5, **1.0** }	{0.5, **1.0**}

B.1 Visual Representation

Model. For contrastive learning on images, we adopt SimCLR [4] as our baseline and follow the same experimental setup as [5]. Specifically, we use the ResNet-50 network as the backbone. To ensure a fair comparison, we set the embedded dimension to 2048 (the representation used in linear readout) and project it into a 128-dimensional space (the actual embedding used for contrastive learning). Regarding the temperature parameter τ, we use the default value τ_0 of 0.5 in most researches, and we also perform grid search on τ varying from 0.1 to 1.0 at an interval of 0.1, denoted by τ^*. The best parameters for each dataset is reported in Table 6. Note that $\{\cdot\}$ indicates the range of hyperparameters that we tune and the numbers in **bold** are the final settings. For α-CL, we follow the setting of [14], where $p = 4$ and $\tau = 0.5$. We use the Adam optimizer with a learning rate of 0.001 and weight decay of $1e - 6$. All models are trained for 400 epochs.

Noisy experiments in Sec.3.4. To investigate the relationship between the temperature parameter τ (or η) and the noise ratio, we follow the approach outlined in [5] and utilize the class information of each image to select negative samples as a combination of true negative samples and false negative samples. Specifically, $r_{ratio} = 0$ indicates all negative samples are true negative samples, $r_{ratio} = 0.5$ suggests 50% of true positive samples existing in negative samples, $r_{ratio} = 1$ means uniform sampling.

Variance analysis in Sec.3.4. To verify the mean-variance objective of InfoNCE, we adopt the approach outlined in [16] and record the negative prediction scores for 256 samples (assuming a batch size of 256) in each minibatch. Specifically, we randomly select samples from a batch to calculate the statistics and visualize them. (1) For positive samples, we calculate cosine similarity by taking the inner product after normalization, and retain the mean value of the 256 positive scores as '*pos mean*'. (2) For negative samples, we average the means and variances of 256 negative samples to show the statistical characteristics of these N negative samples '*(mean neg; var neg)*'. We record this data at each training step to track score distribution throughout the training process.

B.2 Sentence Representation

For the sentence contrastive learning, we adopt the approach outlined in [7] and evaluate our method on 7 popular STS datasets: STS tasks from 2012-2016, STS-B and SICK-R. We utilize the SentEval toolkit to obtain all 7 datasets. Each dataset includes sentence pairs which are rated

Table 7: hyperparameters setting on sentence CL. Note that $\{\cdot\}$ indicates the range of hyperparameters that we tune and the numbers in **bold** are the final settings.

DATASETS	SIMCSE-BERT$_{\text{BASE}}$	SIMCSE-ROBERTA$_{\text{BASE}}$
BEST τ	{ 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, **0.07**, 0.08, 0.09, 0.10, 0.15, 0.20}	{ 0.01, 0.02, 0.03, 0.04, 0.05, **0.06**, 0.07, 0.08, 0.09, 0.10, 0.15, 0.20}
μ	{0.3, **0.4**, 0.5, 0.6, 0.7, 0.8, 0.9 }	{ 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, **2.0**, 2.5, 3.0}
σ	{0.5, **1.0** }	{0.5, **1.0** }

on a scale of 0 to 5, indicating the degree of semantic similarity. To validate the effective of our proposed method, we utilize several methods as baselines: average GloVe embeddings, BERT-flow, BERT-whitening, CT-BERT and SimCSE. The best parameters for each dataset is reported in Table 7. To ensure fairness, we employed the official code, which can be accessed at `https://github.com/princeton-nlp/SimCSE`.

B.3 Graph Representation

For the graph contrastive learning experiments on TU-Dataset [12], we adopted the same experimental setup as outlined in [15]. The dataset statistics can be found in Tab.8. To ensure fairness, we employed the official code, which can be accessed at `https://github.com/Shen-Lab/GraphCL/tree/master/unsupervised_TU`. We made only modifications to the script by incorporating our ADNCE method and conducting experiments on the hyper-parameter $\mu \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $\sigma = 1$ on most datasets. Each parameter was repeated from scratch five times, and the best parameter was selected by evaluating on the validation dataset. The best parameters for each dataset is reported in Table 9.

We summarize the statistics of TU-datasets [12] for unsupervised learning in Table 8. Tab. 10 demonstrates the consistent superiority of our proposed ADNCE approach.

Table 8: Statistics for unsupervised learning TU-datasets.

DATASETS	CATEGORY	GRAPHS#	AVG. N#	AVG. DEGREE
NCI1	BIOCHEMICAL MOLECULES	4,110	29.87	1.08
PROTEINS	BIOCHEMICAL MOLECULES	1,113	39.06	1.86
DD	BIOCHEMICAL MOLECULES	1,178	284.32	715.66
MUTAG	BIOCHEMICAL MOLECULES	188	17.93	19.79
COLLAB	SOCIAL NETWORKS	5,000	74.49	32.99
RDT-B	SOCIAL NETWORKS	2,000	429.63	1.15
RDT-M	SOCIAL NETWORKS	2,000	429.63	497.75
IMDB-B	SOCIAL NETWORKS	1,000	19.77	96.53

Table 9: hyperparameters setting on graph CL. Note that $\{\cdot\}$ indicates the range of hyperparameters that we tune and the numbers in **bold** are the final settings.

DATASETS	BEST τ	μ	σ
NCI1	{ **0.05**, 0.10, 0.15, 0.20, 0.25}	{ 0.2, 0.3 ,0.4, 0.5, 0.6, **0.7**, 0.8, 0.9 }	{ 0.5, **1.0** }
PROTEINS	{ **0.05**, 0.10, 0.15, 0.20, 0.25}	{ 0.5, 1.0, **1.5**, 2.0 }	{ 0.5, **1.0** }
DD	{ 0.05, 0.10, 0.15, **0.20**, 0.25}	{ **0.2**, 0.3 ,0.4, 0.5, 0.6, 0.7, 0.8, 0.9 }	{ 0.5, **1.0** }
MUTAG	{ 0.05, 0.10, **0.15**, 0.20, 0.25}	{ 0.2, 0.3 ,0.4, 0.5, 0.6, **0.7**, 0.8, 0.9 }	{ 0.5, **1.0** }
COLLAB	{ 0.05, **0.10**, 0.15, 0.20, 0.25}	{ **0.2**, 0.3 ,0.4, 0.5, 0.6, 0.7, 0.8, 0.9 }	{ 0.5, **1.0** }
RDT-B	{ 0.05, 0.10, **0.15**, 0.20, 0.25}	{ 0.2, **0.3** ,0.4, 0.5, 0.6, 0.7, 0.8, 0.9 }	{ 0.5, **1.0** }
RDT-M	{ 0.05, 0.10, **0.15**, 0.20, 0.25}	{ 0.2, 0.3 ,0.4, 0.5, 0.6, 0.7, **0.8**, 0.9 }	{ 0.5, **1.0** }
IMDB-B	{ 0.10, 0.20, 0.30, 0.40, **0.50**}	{ 0.2, 0.3 ,0.4, 0.5, 0.6, **0.7**, 0.8, 0.9 }	{ 0.5, **1.0** }

Table 10: Unsupervised representation learning classification accuracy (%) on TU datasets. The compared numbers are from except AD-GCL, whose statistics are reproduced on our platform. **Bold** indicates the best performance while underline indicates the second best on each dataset.

DATASET	NCI1	PROTEINS	DD	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B	AVG.
NO PRE-TRAIN	65.40±0.17	72.73±0.51	75.67±0.29	87.39±1.09	65.29±0.16	76.86 ±0.25	48.48±0.28	69.37±0.37	70.15
INFOGRAPH	76.20± 1.06	74.44± 0.31	72.85± 1.78	89.01±1.13	70.05±1.13	82.50±1.42	53.46±1.03	73.03±0.87	74.02
GRAPHCL	77.87±0.41	74.39±0.45	78.62±0.40	86.80±1.34	71.36±1.15	89.53±0.84	55.99±0.28	71.14±0.44	75.71
AD-GCL	73.91±0.77	73.28±0.46	75.79±0.87	88.74±1.85	72.02±0.56	90.07±0.85	54.33±0.32	70.21±0.68	74.79
RGCL	78.14±1.08	75.03±0.43	78.86±0.48	87.66±1.01	70.92±0.65	90.34±0.58	**56.38±0.40**	**71.85±0.84**	76.15
ADNCE	**79.30±0.67**	**75.10±0.25**	**79.23±0.59**	**89.04±1.30**	**72.26±1.10**	**91.39±0.31**	56.01±0.35	71.58±0.72	**76.74**

References

[1] Aharon Ben-Tal and Marc Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32(11):1445–1466, 1986.

[2] Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

[3] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[6] Louis Faury, Ugo Tanielian, Elvis Dohmatob, Elena Smirnova, and Flavian Vasile. Distributionally robust counterfactual risk minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3850–3857, 2020.

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[8] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4):448–452, 2018.

[9] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2018.

[11] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

[12] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *CoRR*, abs/2007.08663, 2020.

[13] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[14] Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. In *Advances in Neural Information Processing Systems*, 2022.

[15] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020.

[16] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10306–10315, 2021.