Supplementary Material for "About contrastive unsupervised representation learning for classification and its convergence"

A PROOFS FOR SECTION 3

Apart from the similarity between the unsupervised and supervised loss, the proof of Lemma 3.1 uses properties of log-sum-exp.

Proof of Lemma 3.1. We first rewrite the unsupervised loss as:

$$L_{\text{un}}(f) = \mathbb{E}_{(x,x^+)\sim\mathcal{D}_{\text{sim}},x^-\sim\mathcal{D}_{\text{neg}}} \log\left(1+\exp\left(f(x)^T(f(x^-)-f(x^+))\right)\right)$$

where we recognize the ζ function $\zeta(x) = \log(1+e^x)$. We start by using Jensen's inequality

$$\begin{aligned}
L_{\text{un}}(f) &= \mathbb{E}_{\substack{(x,x^+)\sim\mathcal{D}_{\text{sim}}\\ x^-\sim\mathcal{D}_{\text{neg}}}}\left[\zeta\left(f(x)^T(f(x^-)-f(x^+))\right)\right] \\
&\geq \mathbb{E}_{c,c^-\sim\rho,x\sim\mathcal{D}_c}\left[\zeta\left(f(x)^T(\mu_{c^-}-\mu_c)\right)\right] \\
&\geq p_{\min}^{\rho}\mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c}\left[\max_{c^-}\zeta\left(f(x)^T(\mu_{c^-}-\mu_c)\right)\right] \\
&= p_{\min}^{\rho}\mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c}\left[\max_{c^-}\text{LSE}\left(0, f(x)^T(\mu_{c^-}-\mu_c)\right)\right] \\
&\geq p_{\min}^{\rho}\left(\mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c}\left[\text{LSE}\left(f(x)^T(\mu_{c_1}-\mu_c),\dots,f(x)^T(\mu_{c_{N_{\mathcal{C}}}}-\mu_c)\right)\right]-\log N_{\mathcal{C}}\right) \\
&= p_{\min}^{\rho}\left(L_{\text{sup}}^{\mu}(f,\mathcal{C})-\log N_{\mathcal{C}}\right)
\end{aligned}$$

where we have used properties of the log-sum-exp function

$$\max(x_1,\dots,x_n) \leq \text{LSE}(x_1,\dots,x_n) \leq \max(x_1,\dots,x_n)+\log n,$$

the fact that LSE is non-negative whenever one of its arguments is, and for $x\in\mathbb{R}^{2n}$ we have

$$\text{LSE}(x) = \text{LSE}(\text{LSE}(x_1,x_2),\dots,\text{LSE}(x_{2n-1},x_{2n})) \leq \max_{j=1,\dots,n}\text{LSE}(x_{2j-1},x_{2j})+\log n.$$

□

The proof of Lemma 3.2 considers the sample draws where all classes are represented.

Proof of Lemma 3.2. Let $I\in[N_{\mathcal{C}}]^N$ the random vector of classes for each negative sample ($I\sim\rho^{\otimes N}$) and let J be the set of represented classes i.e. $J=\{I_j \mid j\in[N]\}$. We have, again with Jensen's inequality

$$\begin{aligned}
L_{\text{un}}^N(f) &= \mathbb{E}_{x,x^+,x_1^-,\dots,x_N^-}\left[\text{LSE}\left(0, f(x)^T(f(x_1^-)-f(x^+)),\dots,f(x)^T(f(x_N^-)-f(x^+))\right)\right] \\
&\geq \mathbb{E}_{c\sim\rho,I\sim\rho^{\otimes N},x\sim\mathcal{D}_c}\left[\text{LSE}\left(0, f(x)^T(\mu_{I_1}-\mu_c),\dots,f(x)^T(\mu_{I_N}-\mu_c)\right)\right] \\
&\geq \mathbb{P}\left(|J|=N_{\mathcal{C}}\right)\mathbb{E}_{\substack{c\sim\rho\\ I\sim\rho^{\otimes N}\\ x\sim\mathcal{D}_c}}\left[\text{LSE}\left(0, f(x)^T(\mu_{I_1}-\mu_c),\dots,f(x)^T(\mu_{I_N}-\mu_c)\right) \mid |J|=N_{\mathcal{C}}\right] \\
&\geq \mathbb{P}\left(|J|=N_{\mathcal{C}}\right)L_{\text{sup}}^{\mu}(f,\mathcal{C}),
\end{aligned}$$

where we used that for $\mathcal{S}\subset[n]$ and $x\in\mathbb{R}^n$ we have $\text{LSE}(x_{\mathcal{S}})\leq\text{LSE}(x)$ with $x_{\mathcal{S}}$ the restriction of x to the indices in $\mathcal{S}$. Finally, we have $\mathbb{P}\left(|J|=N_{\mathcal{C}}\right)=p_{cc}^{\rho}(N)$. □

We restate Proposition 3.3 for cases $N=1$ and $N>1$. The proof uses Jensen's inequality and the uniformity of ρ.

Proposition 3.3 (restated). *Consider the unsupervised loss $L_{\text{un}}^N(f)$ from Equation (6) with N negative samples. Assume that ρ is uniform over $\mathcal{C}$ and that $2 \leq k+1 \leq N_{\mathcal{C}}$. Then,*

(1) any encoder function $f : \mathcal{X} \to \mathbb{R}^d$ satisfies

$$L_{\text{sup},k}(f) \leq L_{\text{sup},k}^{\mu}(f) \leq \frac{k}{1-\tau^+}\left(L_{\text{un}}(f) - \tau^+\right)$$

with $\tau^+ = \mathbb{P}_{c,c' \sim \rho^2}\left(c = c'\right)$, where $L_{\text{un}}(f)$ is the unsupervised loss from Equation (6) with $N = 1$ negative sample;

(2) more generally,

$$L_{\text{sup},k}(f) \leq L_{\text{sup},k}^{\mu}(f) \leq \frac{k}{1-\tau_N^+}\left(L_{\text{un}}^N(f) - \tau_N^+ \log(N+1)\right)$$

with $\tau_N^+ = \mathbb{P}(c_i = c, \forall i \mid c \sim \rho, (c_1, \cdots, c_N) \sim \rho^N)$, and where $L_{\text{un}}^N(f)$ is the unsupervised loss from Equation (6).

Proof of Proposition 3.3. Let's start with (1). By Jensen's inequality, then use $\log = \log_2$ without loss of generality, and split the expectation into cases $c^- \neq c$ and $c^- = c$,

$$\begin{aligned}
L_{\text{un}}(f) &= \mathbb{E}_{(c,c^-)\sim\rho^2}\mathbb{E}_{x,x^+\sim\mathcal{D}_c,x^-\sim\mathcal{D}_{c^-}}\left[\log\left(1+\exp\left(f(x)^T\left(f(x^-)-f(x^+)\right)\right)\right)\right] \\
&\geq \mathbb{E}_{(c,c^-)\sim\rho^2,x\sim\mathcal{D}_c}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\right] \\
&= (1-\tau^+)\mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c}\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle| c^- \neq c\right] + \tau^+.
\end{aligned}$$

Let us write explicitly the uniform distribution ρ on $\mathcal{C}$. On the one hand,

$$\begin{aligned}
&\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle| c^- \neq c\right] \\
&= \sum_{c^-\in\mathcal{C}\setminus\{c\}} \frac{1}{N_{\mathcal{C}}-1}\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right),
\end{aligned}$$

on the other hand,

$$\begin{aligned}
&\mathbb{E}_{c_1,\ldots,c_k\sim\rho^{\otimes k}}\left[\sum_{i=1}^k \log\left(1+\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right)\middle| \{c,c_1,\ldots,c_k\} \text{ distinct}\right] \\
&= \sum_{\substack{\{c_1,\cdots,c_k\}\subseteq\mathcal{C}\setminus\{c\} \\ \{c_1,\ldots,c_k\}\text{distinct}}} \frac{1}{\binom{N_{\mathcal{C}}-1}{k}} \sum_{i=1}^k \log\left(1+\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right) \\
&= \frac{1}{\binom{N_{\mathcal{C}}-1}{k}} \sum_{\substack{\{c_1,\cdots,c_k\}\subseteq\mathcal{C}\setminus\{c\} \\ \{c_1,\ldots,c_k\}\text{distinct}}} \sum_{i=1}^k \log\left(1+\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right).
\end{aligned}$$

Consider a particular latent class $c^- \in \mathcal{C}\setminus\{c\}$, the term on c^- appears in the double sum for exactly $\binom{N_{\mathcal{C}}-2}{k-1}$ times. And this is for every $c^- \in \mathcal{C}\setminus\{c\}$. We rearrange the double sum according to c^-

$$\begin{aligned}
&= \frac{1}{\binom{N_{\mathcal{C}}-1}{k}}\binom{N_{\mathcal{C}}-2}{k-1} \sum_{c^-\in\mathcal{C}\setminus\{c\}} \log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right) \\
&= k \sum_{c^-\in\mathcal{C}\setminus\{c\}} \frac{1}{N_{\mathcal{C}}-1}\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right).
\end{aligned}$$

Hence, using the uniformity of ρ,

$$\begin{aligned}
&\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|c^-\neq c\right]\\
=&\frac{1}{k}\mathbb{E}_{c_1,\ldots,c_k\sim\rho^{\otimes k}}\left[\sum_{i=1}^{k}\log\left(1+\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right)\middle|\{c,c_1,\ldots,c_k\}\text{ distinct}\right]\\
\geq&\frac{1}{k}\mathbb{E}_{c_1,\ldots,c_k\sim\rho^{\otimes k}}\left[\log\left(1+\sum_{i=1}^{k}\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right)\middle|\{c,c_1,\ldots,c_k\}\text{ distinct}\right].
\end{aligned}$$

That means we have

$$\begin{aligned}
L_{\text{un}}(f)&\geq\frac{1-\tau^+}{k}\mathbb{E}_{\substack{c\sim\rho,x\sim\mathcal{D}_c\\c_1,\ldots,c_k\sim\rho^{\otimes k}}}\left[\log\left(1+\sum_{i=1}^{k}\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right)\middle|\{c,c_1,\ldots,c_k\}\text{ distinct}\right]+\tau^+\\
&=\frac{1-\tau^+}{k}\mathbb{E}_{\mathcal{T}\sim\mathcal{D}^{k+1}}\mathbb{E}_{(x,c)\sim\mathcal{D}_{\mathcal{T}}}\left[-\log\left(\frac{\exp(f(x)^T\mu_c)}{\exp(f(x)^T\mu_c)+\sum_{\substack{c^-\in\mathcal{T}\\c^-\neq c}}\exp\left(f(x)^T\mu_{c^-}\right)}\right)\right]+\tau^+\\
&=\frac{1-\tau^+}{k}L^{\mu}_{\text{sup},k}(f)+\tau^+.
\end{aligned}$$

As for (2), again by Jensen's inequality, and split the expectation into cases $c_i^-=c,\forall i$ and $\exists c_i^-\neq c$,

$$\begin{aligned}
L^N_{\text{un}}(f)&=\mathbb{E}_{(c,c_i^-)\sim\rho^{N+1}}\mathbb{E}_{x,x^+\sim\mathcal{D}_c,x_i^-\sim\mathcal{D}_{c_i^-}}\left[\log\left(1+\sum_{i=1}^{N}\exp\left(f(x)^T\left(f(x_i^-)-f(x^+)\right)\right)\right)\right]\\
&\geq\mathbb{E}_{(c,c_i^-)\sim\rho^{N+1},x\sim\mathcal{D}_c}\left[\log\left(1+\sum_{i=1}^{N}\exp\left(f(x)^T\left(\mu_{c_i^-}-\mu_c\right)\right)\right)\right]\\
&=(1-\tau_N^+)\mathbb{E}_{\substack{c\sim\rho\\x\sim\mathcal{D}_c}}\mathbb{E}_{c_i^-\sim\rho^N}\left[\log\left(1+\sum_{i=1}^{N}\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|\exists c_i^-\neq c\right]+\tau_N^+\log(N+1)
\end{aligned}$$

with

$$\tau_N^+=\mathbb{P}(c_i=c,\forall i\mid c\sim\rho,c_i\sim\rho^N)=\sum_{c\in\mathcal{C}}\rho(c)^{N+1}=N_{\mathcal{C}}^{-N}.$$

Considering the fact that

$$\begin{aligned}
&\mathbb{E}_{c_i^-\sim\rho^N}\left[\log\left(1+\sum_{i=1}^{N}\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|\exists c_i^-\neq c\right]\geq\\
&\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|c^-\neq c\right],
\end{aligned}$$

then by similar computations as in (1), we have

$$L^N_{\text{un}}(f)\geq\frac{1-\tau_N^+}{k}L^{\mu}_{\text{sup},k}(f)+\tau_N^+\log(N+1).$$

□

B Proofs for Section 4

Let us first prove that under Assumption 2, the objective is gradient-Lipschitz w.r.t. the network outputs.

Lemma 4.1. *Consider the unsupervised loss ℓ given by (11), grant Assumption 2 and define the set*

$$B^3=\left\{z=(z_1,z_2,z_3)\in(\mathbb{R}^d)^3\ :\ \max_{j=1,2,3}\|z_j\|_2^2\leq C^2\right\}$$

where $C>0$ is defined in Assumption 2. Then, the restriction of ℓ to B^3 satisfies (12) with a constant $L_{\text{smooth}}\leq 2+8C^2$.

Proof. We will prove this result by bounding the norm of the Hessian matrix.

Let us write the gradient of $\ell(z)$ with respect to z first. We have $z \in \mathbb{R}^{3d}$. For ease of writing, we define the matrices $A_1, A_2, A_3 \in \mathbb{R}^{3d \times d}$ as

$$A_1 = \begin{pmatrix} I_d \\ 0_d \\ 0_d \end{pmatrix} \quad A_2 = \begin{pmatrix} 0_d \\ I_d \\ 0_d \end{pmatrix} \quad A_3 = \begin{pmatrix} 0_d \\ 0_d \\ I_d \end{pmatrix}$$

where $I_d, 0_d \in \mathbb{R}^{d \times d}$ are the identity and zero matrix respectively. With this notation, we have $z_i = A_i^T z$ for $i = 1, 2, 3$ the three contiguous thirds of z's coordinates.

Our purpose is to compute

$$\frac{\partial}{\partial z}\ell(z) = \frac{\partial}{\partial z}\left[-\log\left(\frac{\exp\left(z_1^T z_2\right)}{\exp\left(z_1^T z_2\right) + \exp\left(z_1^T z_3\right)}\right)\right].$$

Denote $\cos_{i,j} = z_i^T z_j$, we can now compute for $i, j \in \{1, 2, 3\}$ $(i \neq j)$

$$\frac{\partial}{\partial z}\cos_{i,j} = \left(A_i A_j^T + A_j A_i^T\right) z =: \partial \cos_{i,j} \in \mathbb{R}^{3d}.$$

Now, denote $v = \operatorname{softmax}\left(\cos_{1,2}, \cos_{1,3}\right) \in \mathbb{R}^2$, we can write

$$\frac{\partial}{\partial z}\ell(z) = (v_1 - 1)\partial \cos_{1,2} + v_2 \partial \cos_{1,3}\,.$$

We proceed with the following computation

$$\frac{\partial^2}{\partial z^2}\cos_{i,j} = A_i A_j^T + A_j A_i^T,$$

which we will denote simply as $\partial^2 \cos_{i,j}$. Before we get the Hessian of loss, we still need to compute

$$\partial v := \frac{\partial v}{\partial z} = \left(\operatorname{diag}(v) - vv^T\right)\begin{pmatrix} \partial \cos_{1,2}^T \\ \partial \cos_{1,3}^T \end{pmatrix} \in \mathbb{R}^{2 \times 3d}.$$

Now we can write

$$\frac{\partial^2}{\partial z^2}\ell(z) = (v_1 - 1)\partial^2 \cos_{1,2} + v_2 \partial^2 \cos_{1,3} + \begin{pmatrix} \partial \cos_{1,2} & \partial \cos_{1,3} \end{pmatrix} \partial v.$$

We can now estimate the norm of this matrix which will provide an estimation for the Lipschitz constant.

We find that

$$\left\|\partial \cos_{i,j}\right\| \leq 2 \max\left(\|z_i\|, \|z_j\|\right),$$

keeping in mind that the matrix $\operatorname{diag}(v) - vv^T$ has norm at most $1/2$, this leads to

$$\left\|\begin{pmatrix} \partial \cos_{1,2} & \partial \cos_{1,3} \end{pmatrix} \partial v\right\| = 8 \max_{i,j}\left(\|z_i\| \|z_j\|\right).$$

We have also that $\left\|\partial^2 \cos_{i,j}\right\| = 1$.

All in all, we have $\left\|\frac{\partial^2}{\partial z^2}\ell(z)\right\| = 2 + 8 \max_{i,j}\left(\|z_i\| \|z_j\|\right)$. Recalling that we restricted $\mathbb{R}^{3d}$ so that we have $\max_i \|z_i\| \leq C$ the result follows. □

Theorem 1 is actually obtained in two steps. First, Theorem 6 from Allen-Zhu et al. (2019) allows us to obtain that the gradient of the objective $\nabla \widehat{L}_{\text{un}}(f)$ with respect to the network outputs reaches arbitrarily low values. Then, combining this with Assumption 2, this result can be extended into the objective itself.

Following appendix A of Allen-Zhu et al. (2019), we need to define the loss vectors for our model. These are originally defined as $\text{loss}_i = y_i - y_i^*$ (y_i and y_i^* are respectively the output and label corresponding to an input x_i from the dataset) for the ℓ^2 loss. More generally, for a network output z_i, they are defined as

$$\text{loss}_i = \nabla_z \ell(z_i).$$

Following the unsupervised training protocol, samples are fed into the network three at a time x, x^+ and x^-. Let us denote θ the parameters of the network f, for a triplet (x_i, x_i^+, x_i^-), the trick is to write:

$$\frac{\partial}{\partial \theta}\ell(z_i) = \frac{\partial z}{\partial \theta}\underbrace{\frac{\partial}{\partial z}\ell(z_i)}_{\text{loss}}$$

with z_i the concatenation of $f(x_i), f(x_i^+), f(x_i^-)$.

By denoting $(x_1, x_2, x_3) = (x_i, x_i^+, x_i^-)$, the previous writing is equivalent to

$$\sum_{j=1}^{3} \frac{\partial f(x_j)}{\partial \theta} A_j^T \frac{\partial}{\partial z}\ell(z_i)$$

and by letting $\text{loss}_{i,j} = A_j^T \frac{\partial}{\partial z}\ell(z_i)$, we obtain a triplet of loss vectors for each data triple (matrices A_j defined in the previous proof).

Lemma B.1. *Grant Assumption 1 and let $\widehat{L}_{\text{un}}(f)$ be the loss incurred by f:*

$$\widehat{L}_{\text{un}}(f) = \sum_{i=1}^{n} \ell(f(x_i), f(x_i^+), f(x_i^-))$$

and let $\epsilon > 0$ be the desired precision. Then, assuming $m \geq \Omega\left(\text{poly}(n, L, \delta^{-1}) \cdot d\epsilon^{-1}\right)$, the gradient descent with learning rate $\nu = \Theta\left(\frac{d\delta}{\text{poly}(n,L)\cdot m}\right)$ finds parameters such that

$$\|\nabla \widehat{L}_{\text{un}}(f)\| \leq \epsilon$$

after a number of steps $T = O\left(\frac{\text{poly}(n,L)}{\delta^2\epsilon^2}\right)$.

Proof. This result follows from Allen-Zhu et al. (2019) (see Theorem 6 and appendix A). It corresponds to the case of a non-convex bounded loss function. We only need to check the used loss function ℓ is bounded and gradient-Lipschitz smooth. The latter condition is verified due to Lemma 4.1 and Assumption 2.

As for the boundedness, it is also a consequence of Assumption 2 and the fact that the softplus function satisfies

$$\zeta(x) \sim^{x \to +\infty} x \quad \text{and} \lim_{x \to -\infty} \zeta(x) = 0.$$

□

From here, we can derive a result for the objective itself (Theorem 1) thanks to the following Lemma.

Lemma 4.2. *Grant Assumption 2 and assume that the parameters of the encoder f are optimized so that $\|\nabla \widehat{L}_{\text{un}}(f)\| \leq \epsilon$ with $\epsilon < \eta/2$, where η is defined in Assumption 2. Then, for any $i = 1, \ldots, n$, we have $\ell(z_i) \leq 2\epsilon/\eta$ where $z_i = (f(x_i), f(x_i^+), f(x_i^-))$.*

Proof. Since we assume $\|\nabla \widehat{L}_{\text{un}}(f)\| \leq \epsilon$, this also implies that $\max_{i,j} \|\text{loss}_{i,j}\| \leq \epsilon$ (see Theorem 3 of Allen-Zhu et al. (2019) and its variant in appendix A).

We can write the norms $\|\text{loss}_{i,j}\|$ as:

$$\begin{aligned}
\|\text{loss}_{i,1}\| &= \|(v_1 - 1)z_{i,2} + v_2 z_{i,3}\| \\
\|\text{loss}_{i,2}\| &= |v_1 - 1| \, \|z_{i,1}\| \\
\|\text{loss}_{i,3}\| &= v_2 \, \|z_{i,1}\|
\end{aligned}$$

where we defined $v = \mathrm{softmax}(z_1^T z_2, z_1^T z_3)$.

Thanks to Assumption 2, we can argue that $\|z_{i,j}\| \geq \eta$. These quantities can be small for $v_1 \to 1$ and $v_2 \to 0$. Since we have $\max_{i,j} \|\mathrm{loss}_{i,j}\| \leq \epsilon$, this implies in particular that for all i we get $\|\mathrm{loss}_{i,3}\| \leq \epsilon$ which means $v_2 \leq \epsilon/\eta$, and we have $v_2 = \sigma(z_1^T(z_3 - z_2))$. So for an instance $i \in [n]$ the loss term in the objective is:

$$\begin{aligned}
\zeta(z_{i,1}^T(z_{i,3} - z_{i,2})) &= \log(1 + \exp(z_{i,1}^T(z_{i,3} - z_{i,2}))) \\
&= -\log(\sigma(-z_{i,1}^T(z_{i,3} - z_{i,2}))) \\
&= -\log(1 - \sigma(z_{i,1}^T(z_{i,3} - z_{i,2}))) \\
&= -\log(1 - v_2) \leq \frac{v_2}{1 - v_2} \leq 2v_2 \leq 2\epsilon/\eta,
\end{aligned}$$

where we used the inequality $-\log(1 - x) \leq \frac{x}{1-x}$ for $0 \leq x < 1$, and the assumption that $\epsilon < \eta/2$. □

Lemma 4.2 allows us to deduce that the objective is well optimized (we treated the loss term for a single triplet here but the same methods can be applied to the whole objective with a number of gradient steps which is still polynomial).

Proof of Theorem 1. Theorem 1 is the consequence of combining Lemma B.1 applied using $\frac{\epsilon\eta}{2n}$ instead of ϵ and Lemma 4.2 (the $1/n$ factor can be absorbed by the $\mathrm{poly}(n, L)$ factors in the bounds of Lemma B.1). □