Supplementary Material: Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation

A Detailed Proofs

A.1 Proof of Corollary 1

Proof. Removing the supremum and rearranging the terms in Lemma 1, we prove our claim. □

A.2 Proof of Lemma 2

Proof. By Theorem 1, we want to find

$$\begin{aligned}(\mathbb{I}^R_{f,P})^*(\phi) &= \sup_{p\in\Delta(P)} \int_{\mathcal{H}} \phi p - (p-1)^2 dP \\ &= \sup_{p\in\Delta(P)} \mathbb{E}_P[\phi p - (p-1)^2]\end{aligned} \tag{6}$$

where ϕ is a measurable function on $\mathcal{H}$. In order to find the supremum on the right hand side, we consider the following Lagrangian:

$$\begin{aligned}L(p,\lambda) &= \mathbb{E}_P[\phi p - (p-1)^2] + \lambda(\mathbb{E}_P[p]-1) \\ &= \int_{\mathcal{H}} \phi p - (p-1)^2 dP + \lambda\bigg(\int_{\mathcal{H}} p dP - 1\bigg)\end{aligned}$$

where $\lambda \in \mathbb{R}$ and p is constrained to be a probability density over $\mathcal{H}$. By talking the functional derivative with respect to p and renormalize by dropping the dP factor multiplying all terms and setting it to zero, we have $\frac{\partial L}{\partial p dP} = \phi - 2(p-1) + \lambda = 0$. Thus, we have $p = \frac{\phi+\lambda}{2} + 1$. Since p is constrained to be $\int_{\mathcal{H}} p dP = 1$, we have $\int_{\mathcal{H}}(\frac{\phi+\lambda}{2}+1)dP = 1$. Then, we obtain $\lambda = -\mathbb{E}_P[\phi]$ and the optimum $p = \frac{\phi - \mathbb{E}_P[\phi]}{2} + 1$. Plugging it in Equation (6), we have

$$(\mathbb{I}^R_{f,P})^*(\phi) = E_P\bigg[\phi\bigg(\frac{\phi - \mathbb{E}_P[\phi]}{2} + 1\bigg) - \bigg(\frac{\phi - \mathbb{E}_P[\phi]}{2}\bigg)^2\bigg]$$

which simplifies to $\mathbb{E}_P[\phi] + \frac{1}{4}\mathbb{V}\mathrm{ar}_P[\phi]$. □

A.3 Proof of Lemma 3

Proof. Notice that the χ^2-divergence is obtained by setting $f(x) = (x-1)^2$. In order to find the convex conjugate $f^*(y)$ from Equation (1), let $g(x,y) = xy - (x-1)^2$. We need to find the supremum of $g(x,y)$ with respect to x. By using differentiation with respect to x and setting the derivative to zero, we have $\frac{\partial g}{\partial x} = y - 2(x-1) = 0$. Thus, plugging $x = \frac{y}{2} + 1$ for $g(x,y) = xy - (x-1)^2$, we obtain $f^*(y) = y + \frac{y^2}{4}$. Plugging $f(x)$ and $f^*(y)$ in Corollary 1, we prove our claim. □

A.4 Proof of Lemma 4

Proof. By Theorem 1, we want to find

$$(\mathbb{I}^R_{f,P})^*(\phi) = \sup_{p\in\Delta(P)} \mathbb{E}_P[\phi p - \frac{1}{2}|p-1|]$$

In order to find the supremum on the right hand side, we consider the following Lagrangian

$$\begin{aligned} L(p,\lambda) &= \mathbb{E}_P[\phi p - \frac{1}{2}|p-1|)] + \lambda(\mathbb{E}_P[p]-1) \\ &= \begin{cases} \mathbb{E}_P[\phi p - \frac{1}{2}p + \frac{1}{2}] + \lambda(\mathbb{E}_P[p]-1) & \text{if } 1 \le p \\ \mathbb{E}_P[\phi p + \frac{1}{2}p - \frac{1}{2}] + \lambda(\mathbb{E}_P[p]-1) & \text{if } 0 < p < 1 \end{cases} \end{aligned}$$

Then, it is not hard to see that if $|\phi| \le \frac{1}{2}$, $L(p,\lambda)$ is maximized at $p=1$, otherwise $L \to \infty$ as $p \to 1$. Thus, Lemma 4 holds for $|\phi| \le \frac{1}{2}$. If we add $\frac{1}{2}$ on the both sides, then ϕ is bounded between 0 and 1, as we claimed in the corollary. □

A.5 Proof of Lemma 5

Proof. The corresponding convex function $f(x)$ for the α-divergence is defined as $f(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$. Applying the same procedure as in Lemma 3, we get the convex conjugate $f^*(y) = \frac{(\alpha-1)^{\frac{\alpha}{\alpha-1}}}{\alpha} y^{\frac{\alpha}{\alpha-1}} + \frac{1}{\alpha(\alpha-1)}$ for $\alpha > 1$. Plugging $f(x)$ and $f^*(y)$ in Corollary 1, we prove our claim. □

A.6 Proof of Lemma 6

Proof. The corresponding convex function $f(x)$ for the squared Hellinger divergence is defined as $f(x) = (\sqrt{x}-1)^2$. Let $g(x,y) = xy - (\sqrt{x}-1)^2$. We have $\frac{\partial g}{\partial x} = y - 1 + \frac{1}{\sqrt{x}} = 0$. Since we consider $x > 0$, $y < 1$. Applying the same procedure as in Lemma 3, we get the convex conjugate $f^*(y) = \frac{y}{1-y}$. Plugging $f(x)$ and $f^*(y)$ in Corollary 1, we prove our claim. □

A.7 Proof of Lemma 7

Proof. The corresponding convex function $f(x)$ for the reverse KL-divergence is defined as $f(x) = -\log(x)$. Applying the same procedure as in Lemma 6, we get the convex conjugate $f^*(y) = \log(-\frac{1}{y}) - 1$. Plugging $f(x)$ and $f^*(y)$ in Corollary 1, we have

$$\mathbb{E}_Q[\psi] \le \overline{KL}(Q\|P) + \mathbb{E}_P\left[\log\left(-\frac{1}{\psi}\right)\right] - 1$$

where $\psi < 0$. Letting $\psi = \phi - 1$, we prove our claim. □

A.8 Proof of Lemma 8

Proof. The corresponding convex function $f(x)$ for the Neyman χ^2-divergence is defined as $f(x) = -\frac{(x-1)^2}{x}$. Applying the same procedure as in Lemma 6, we get the convex conjugate $f^*(y) = 2 - 2\sqrt{1-y}$. Plugging $f(x)$ and $f^*(y)$ in Corollary 1, we prove our claim. □

A.9 Proof of Lemma 10

Proof.

$$\begin{aligned} \mathbb{E}_Q[\phi] &= \int_{\mathcal{H}} \phi dQ \\ &= \int_{\mathcal{H}} \phi \frac{dQ}{dP} dP \\ &\le \left(\int_{\mathcal{H}} \left|\frac{dQ}{dP}\right|^\alpha dP\right)^{\frac{1}{\alpha}} \left(\int_{\mathcal{H}} |\phi|^{\frac{\alpha}{\alpha-1}} dP\right)^{\frac{\alpha-1}{\alpha}} \\ &= \left(\alpha(\alpha-1)D_\alpha(Q\|P) + 1\right)^{\frac{1}{\alpha}} \left(\mathbb{E}_P[|\phi|^{\frac{\alpha}{\alpha-1}}]\right)^{\frac{\alpha-1}{\alpha}} \end{aligned}$$

The third line is due to the Hölder's inequality. □

A.10 Proof of Lemma 12

Proof. Consider the covariance of ϕ and $\frac{dQ}{dP}$.

$$\begin{aligned}\left|Cov_P\left(\phi, \frac{dQ}{dP}\right)\right| &= \left|\int_{\mathcal{H}} \phi \frac{dQ}{dP} dP - \int_{\mathcal{H}} \phi dP \int_{\mathcal{H}} \frac{dQ}{dP} dP\right| \\ &= \left|\mathbb{E}_Q[\phi] - \mathbb{E}_P[\phi]\right|\end{aligned}$$

On the other hand,

$$\begin{aligned}&\left|Cov_P\left(\phi, \frac{dQ}{dP}\right)\right| \\ &= \left|\mathbb{E}_P\left[(\phi - \mu_P)\left(\frac{dQ}{dP} - \mathbb{E}_P\left[\frac{dQ}{dP}\right]\right)\right]\right| \\ &\leq \mathbb{E}_P\left[\left|(\phi - \mu_P)\left(\frac{dQ}{dP} - \mathbb{E}_P\left[\frac{dQ}{dP}\right]\right)\right|\right] \\ &= \int_{\mathcal{H}} \left|(\phi - \mu_P)\left(\frac{dQ}{dP} - 1\right)\right| dP \\ &\leq \left(\int_{\mathcal{H}} \left|\frac{dQ}{dP} - 1\right|^{\alpha} dP\right)^{\frac{1}{\alpha}} \left(\int_{\mathcal{H}} |\phi - \mu_P|^{\frac{\alpha}{\alpha-1}} dP\right)^{\frac{\alpha-1}{\alpha}} \\ &\leq \widetilde{\mathcal{D}_{\alpha}}(Q\|P)^{\frac{1}{\alpha}} \left(\mathbb{E}_P[|\phi - \mu_P|^{\frac{\alpha}{\alpha-1}}]\right)^{\frac{\alpha-1}{\alpha}}\end{aligned}$$

which proves our claim. □

A.11 Proof of Proposition 2

Proof. Note that we have Lemma 12. Now, it remains to bound $\mathbb{E}_P[\phi(X)]$ and $\mathbb{E}_P[|\phi(X) - \mathbb{E}_P[\phi(X)]|^{\frac{\alpha}{\alpha-1}}]$. Since P is a strongly log-concave distribution and ϕ is an L-Lipschitz function, Theorem 3.16 in Wainwright (2019) implies $\phi(X) - \mathbb{E}_P[\phi(X)]$ is a sub-Gaussian random variable with parameter $\sigma^2 = \frac{2L^2}{\gamma}$. Therefore, we have

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \mathbb{E}_P[\phi(X)]\right| \geq \epsilon\right) \leq 2e^{\frac{n\gamma\epsilon^2}{4L^2}}$$

Thus,

$$\left|\mathbb{E}_P[\phi(X)] - \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\right| \underset{1-\delta}{\leq} \frac{4L^2}{n\gamma} \log(\frac{2}{\delta})$$

On the other hand, letting $U = \phi(X) - \mathbb{E}_P[\phi(X)]$, then we have

$$\begin{aligned}\mathbb{E}_P[|U|^{\frac{\alpha}{\alpha-1}}] &= \int_0^\infty Pr\{|U|^{\frac{\alpha}{\alpha-1}} > u\}du \\ &= \int_0^\infty Pr\{|U| > u^{\frac{\alpha-1}{\alpha}}\}du \\ &= \frac{\alpha}{\alpha-1}\int_0^\infty u^{\frac{1}{\alpha-1}} Pr\{|U| > u\}du \\ &= \frac{\alpha}{\alpha-1}\int_0^\infty u^{\frac{1}{\alpha-1}} Pr\{|\phi(X) - \mathbb{E}_P[\phi(X)]| > u\}du \\ &\leq \frac{\alpha}{\alpha-1}\int_0^\infty u^{\frac{1}{\alpha-1}} 2e^{-\frac{\gamma u^2}{4L^2}}du \\ &= \frac{\alpha}{\alpha-1}\int_0^\infty u^{\frac{\alpha}{2(\alpha-1)}-1} e^{-\frac{\gamma u}{4L^2}}du \\ &= \frac{\alpha}{\alpha-1}\Gamma\Big(\frac{\alpha}{2(\alpha-1)}\Big)\Big(\frac{4L^2}{\gamma}\Big)^{\frac{\alpha}{2(\alpha-1)}} \\ &= 2\Gamma\Big(\frac{3\alpha-2}{2(\alpha-1)}\Big)\Big(\frac{4L^2}{\gamma}\Big)^{\frac{\alpha}{2(\alpha-1)}}\end{aligned}$$

The inequality is due to sub-Gaussianity. Putting everything together with Lemma 12, we prove our claim. □

A.12 Proof of Proposition 3

Proof. By Lemma 11, we have

$$|E_Q[\phi(X)] - \mathbb{E}_P[\phi(X)]| \leq \sqrt{\chi^2(Q\|P)\mathbb{E}_P[\phi^2(X)]}$$

where $\mathbb{E}_P[\phi(X)]$ is bounded as in Proposition 2.

$$\left|\mathbb{E}_P[\phi(X)] - \frac{1}{n}\sum_{i=1}^n \phi(X_i)\right| \underset{1-\delta}{\leq} \frac{4L^2}{n\gamma}\log(\frac{2}{\delta})$$

It remains to bound $E_P[\phi^2(X)]$. Note that U^2 is a sub-exponential random variable with parameters $(4\sqrt{2}\sigma^2, 4\sigma^2)$ when U is a sub-Gaussian random variable with a parameter σ (See Appendix B in Honorio and Jaakkola (2014) for details). Therefore,

$$\begin{aligned}&Pr\left(\left|\frac{1}{n}\sum_{i=1}^n \phi^2(X_i) - \mathbb{E}_P[\phi^2(X)]\right| \geq \epsilon\right) \\ &\qquad\qquad \leq \begin{cases} 2e^{-\frac{n\gamma^2\epsilon^2}{256L^4}} \text{if } \log\frac{2}{\delta} \leq n \\ 2e^{-\frac{n\gamma^2\epsilon^2}{256L^4}} \text{if } \log\frac{2}{\delta} > n \end{cases}\end{aligned}$$

Thus,

$$\begin{aligned}&\left|\mathbb{E}_P[\phi^2(X)] - \frac{1}{n}\sum_{i=1}^n \phi^2(X_i)\right| \\ &\qquad \underset{1-\delta}{\leq} \begin{cases} \frac{16L^2}{n}\sqrt{\frac{1}{n}\log(\frac{2}{\delta})} & \text{if } \log\frac{2}{\delta} \leq n \\ \frac{16L^2}{n\gamma}\log(\frac{2}{\delta}) & \text{if } \log\frac{2}{\delta} > n \end{cases}\end{aligned}$$

Putting everything together, we prove our claim. □

A.13 Proof of Proposition 4

Proof. Note that we have the following change of measure inequality for any function $\psi : \mathbb{R} \rightarrow \mathbb{R}$ from the Donsker-Varadhan representation for the KL-divergence.

$$\mathbb{E}_Q[\psi(X)] \leq KL(Q\|P) + \log(\mathbb{E}_P[e^{\psi(X)}]) \tag{7}$$

Suppose $\psi_1 : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the following condition.

$$\psi_1(X) = \phi(X) - \mathbb{E}_P[\phi(X)]$$

By plugging in Equation (7), we have

$$\begin{aligned}&\mathbb{E}_Q[\phi(X)] - \mathbb{E}_P[\phi(X)]\\ &\leq KL(Q\|P) + \log(\mathbb{E}_P[e^{\phi(X)-\mathbb{E}_P[\phi(X)]}])\end{aligned}$$

On the other hand, suppose $\psi_2 : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the following condition.

$$\psi_2(X) = \mathbb{E}_P[\phi(X)] - \phi(X)$$

By plugging in Equation (7), we have

$$\begin{aligned}&\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]\\ &\leq KL(Q\|P) + \log(\mathbb{E}_P[e^{\mathbb{E}_P[\phi(X)]-\phi(X)}])\end{aligned}$$

By putting all together, we have

$$\begin{aligned}|\mathbb{E}_Q[\phi(X)] - \mathbb{E}_P[\phi(X)]| &\leq KL(Q\|P) +\\ &\log(\mathbb{E}_P[e^{\phi(X)-\mathbb{E}_P[\phi(X)]}]) \vee \log(\mathbb{E}_P[e^{\mathbb{E}_P[\phi(X)]-\phi(X)}])\end{aligned}$$

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables.We can easily see that we have

$$\begin{aligned}&|\mathbb{E}_Q[\phi(X)] - \mathbb{E}_P[\phi(X)]| \leq KL(Q\|P) +\\ &\log(\mathbb{E}_P[e^{\frac{1}{n}\sum_{i=1}^n \phi(X_i)-\mathbb{E}_P[\phi(X)]}]) \vee \log(\mathbb{E}_P[e^{\mathbb{E}_P[\phi(X)]-\frac{1}{n}\sum_{i=1}^n \phi(X_i)}])\end{aligned} \tag{8}$$

$\mathbb{E}_P[\phi(X)]$ is bounded as in Proposition 2.

$$\left|\mathbb{E}_P[\phi(X)] - \frac{1}{n}\sum_{i=1}^n \phi(X_i)\right| \underset{1-\delta}{\leq} \frac{4L^2}{n\gamma}\log(\frac{2}{\delta}) \tag{9}$$

As shown in Proposition 2, $\phi(X) - \mathbb{E}_P[\phi(X)]$ is a sub-Gaussian random variable with parameter $\sigma^2 = \frac{2L^2}{\gamma}$. Therefore, by Definition 2,

$$\log(\mathbb{E}_P[e^{\frac{1}{n}\sum_{i=1}^n \phi(X_i)-\mathbb{E}_P[\phi(X)]}]) \vee \log(\mathbb{E}_P[e^{\mathbb{E}_P[\phi(X)]-\frac{1}{n}\sum_{i=1}^n \phi(X_i)}]) \leq \frac{L^2}{n\gamma} \tag{10}$$

since both values in the maximum are bounded by $\frac{L^2}{n\gamma}$. By (8), (9) and (10), we prove our claim. □

B Pseudo α-divergence is a member of the family of f-divergences.

Proof. Let $f(t) = |t-1|^\alpha$ and $\lambda \in [0,1]$. Let $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ for $\alpha \geq 1$.

$$\begin{aligned}&f(\lambda x + (1-\lambda)y)\\ &= |\lambda x + (1-\lambda)y - 1|^\alpha\\ &= |\lambda(x-1) + (1-\lambda)(y-1)|^\alpha\\ &\leq \{\lambda|x-1| + (1-\lambda)|y-1|\}^\alpha\\ &= \{\lambda^{\frac{1}{\alpha}}|x-1|\lambda^{\frac{1}{\beta}} + (1-\lambda)^{\frac{1}{\alpha}}|y-1|(1-\lambda)^{\frac{1}{\beta}}\}^\alpha\\ &\leq (\lambda|x-1|^\alpha + (1-\lambda)|y-1|^\alpha)(\lambda + (1-\lambda))^{\frac{\alpha}{\beta}}\\ &= \lambda|x-1|^\alpha + (1-\lambda)|y-1|^\alpha\\ &= \lambda f(x) + (1-\lambda)f(y)\end{aligned}$$

The first inequality is due to the triangle inequality and the second inequality is due to Hölder's inequality. By the definition of convexity, $f(t)$ is convex. Also, $f(1) = 0$. By the definition of f-divergence, we prove our claim. □

C PAC-Bayesian bounds for bounded, sub-Gaussian and sub-exponential losses

Here we complement our results in Section 3, by showing our PAC-Bayesian generalization bounds for bounded, sub-Gaussian and sub-exponential losses.

C.1 Bounded Loss Function

First, let us assume here that the loss function is bounded, i.e., for any $h \in \mathcal{H}$ and $(x, y) \in \mathcal{X}\times\mathcal{Y}$, $\ell(h(x), y) \in [0, R]$ for $R > 0$. Note that, for $R > 1$, we cannot use the total variation, squared Hellinger, Reverse KL and Neyman χ^2 divergence because ϕ is constrained to be in $[0, 1]$.

Proposition 5 (The PAC-Bayesian bounds for bounded loss function)**.** *Let P be any prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size $m > 0$ and $\alpha > 1$, with probability at least $1 - \delta$, simultaneously for all posterior distributions Q, we have*

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{R^2}{2m}\log(\frac{2}{\delta})\big(\alpha(\alpha-1)D_\alpha(Q\|P)+1\big)^{\frac{1}{\alpha}}} \tag{11}$$

$$\begin{aligned} &R_D(G_Q) \leq R_S(G_Q)+ \\ &\sqrt{\frac{1}{m}\left(D_\alpha(Q\|P)+\frac{1}{\alpha(\alpha-1)}\right)+\frac{1}{m\alpha}\left(\frac{R^2(\alpha-1)}{2}\log(\frac{2}{\delta})\right)^{\frac{\alpha}{\alpha-1}}} \end{aligned} \tag{12}$$

Proof. Suppose that we have a convex function $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, that measures the discrepancy between the observed empirical Gibbs risk $R_S(G_Q)$ and the true Gibbs risk $R_D(G_Q)$ on distribution Q. Given that, the purpose of the PAC-Bayesian theorem is to upper-bound the discrepancy $t\Delta(R_D(G_Q), R_S(G_Q))$ for any $t > 0$. Let $\phi_D(h) := t\Delta(R_D(h), R_S(h))$, where the subscript of ϕ_D shows the dependency on the data distribution D. Let $\Delta(q, p) = (q - p)^2$. By applying Jensen's inequality on convex function Δ for the first step,

$$\begin{aligned} t\Delta(R_D(G_Q), R_S(G_Q)) &= t\Delta(\mathop{\mathbb{E}}_{h\sim Q} R_D(h), \mathop{\mathbb{E}}_{h\sim Q} R_S(h)) \\ &\leq \mathop{\mathbb{E}}_{h\sim Q} t\Delta(R_D(h), R_S(h)) \\ &= \mathop{\mathbb{E}}_{h\sim Q} \phi_D(h) \end{aligned} \tag{13}$$

where $\phi_D(h) = t\Delta(R_D(h), R_S(h)) = t(R_S(h)-R_D(h))^2$. By Hoeffding's inequality, for any $\epsilon > 0$ and any $h \in \mathcal{H}$,

$$\begin{aligned} &\mathop{Pr}_{(x,y)\sim D}(\phi_D(h) \geq \epsilon) = \mathop{Pr}_{(x,y)\sim D}\left(\big(R_S(h) - R_D(h)\big)^2 \geq \frac{\epsilon}{t}\right) \\ &= \mathop{Pr}_{(x,y)\sim D}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i), y_i) - \mathop{\mathbb{E}}_{(x,y)\sim D}\ell(h(x), y)\right| \geq \sqrt{\frac{\epsilon}{t}}\right) \\ &\leq 2e^{-\frac{2m\epsilon}{R^2 t}} \end{aligned}$$

Setting $\delta = 2e^{-\frac{2m\epsilon}{R^2 t}}$, we have

$$\phi_D(h) \underset{1-\delta}{\leq} \frac{tR^2}{2m}\log(\frac{2}{\delta})$$

The symbol $\underset{1-\delta}{\leq}$ denotes that the inequality holds with probability at least $1 - \delta$. The second line holds due to the Hoeffding's inequality. For $\alpha > 1$, we have

$$(\phi_D(h))^{\frac{\alpha}{\alpha-1}} \underset{1-\delta}{\leq} \left(\frac{tR^2}{2m}\log(\frac{2}{\delta})\right)^{\frac{\alpha}{\alpha-1}}$$

Also note that

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim P}[(\phi_D(h))^{\frac{\alpha}{\alpha-1}}] &\leq \sup_{(x,y)\sim D}\left\{\mathop{\mathbb{E}}_{h\sim P}[(\phi_D(h))^{\frac{\alpha}{\alpha-1}}]\right\}\\ &\leq \mathop{\mathbb{E}}_{h\sim P}\left[\sup_{(x,y)\sim D}(\phi_D(h))^{\frac{\alpha}{\alpha-1}}\right]\\ &\underset{1-\delta}{\leq}\left(\frac{tR^2}{2m}\log(\frac{2}{\delta})\right)^{\frac{\alpha}{\alpha-1}}\end{aligned} \tag{14}$$

By applying Equations (13) and (14) to Lemma 10,

$$t(R_D(G_Q)-R_S(G_Q))^2\underset{1-\delta}{\leq}\frac{tR^2}{2m}\log(\frac{2}{\delta})\Big(\alpha(\alpha-1)D_\alpha(Q\|P)+1\Big)^{\frac{1}{\alpha}}$$

which proves Equation (11). By applying Equations (13) and (14) to Lemma 5 and setting $t=m$, we prove Equation (12). □

Equation (11) has a tighter bound than Proposition 2 in Alquier and Guedj (2018). Also, Equation (12) has a novel expression of PAC-Bayesian theorem with χ^2-divergence. The same arguments apply to the bounds in Proposition 6, 7 and 1.

The following corollary is an immediate consequence of this proposition for $\alpha=2$.

Corollary 3 (The PAC-Bayesian bounds with χ^2-divergence for bounded loss function)**.** *Let P be any prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size $m>0$, with probability at least $1-\delta$, simultaneously for all posterior distributions Q, we have*

$$R_D(G_Q)\leq R_S(G_Q)+\sqrt{\frac{R^2}{2m}\log(\frac{2}{\delta})\sqrt{\chi^2(Q\|P)+1}} \tag{15}$$

$$\begin{aligned}R_D(G_Q)&\leq R_S(G_Q)\\ &+\sqrt{\frac{1}{2m}\left(\chi^2(Q\|P)+1+\left(\frac{R^2}{2}\log(\frac{2}{\delta})\right)^2\right)}\end{aligned} \tag{16}$$

These are novel PAC-Bayesian bounds for bounded loss functions based on χ^2-divergence. Most PAC-Bayesian bounds for bounded loss functions are based on the KL-divergence for the complexity term (see e.g., Catoni (2007); Seeger (2003); McAllester (1998)). Honorio and Jaakkola (2014) and Bégin et al. (2016) contain bounds for bounded loss functions with χ^2-divergence. Compared to their bounds, the bound (15) is tighter due to the power of $\frac{1}{4}$ on the complexity term and $\log(\frac{1}{\delta})$ instead of $\frac{1}{\delta}$. Also, PAC-Bayes bound (16) has a unique characteristics; $\chi^2(Q\|P)$ and $\frac{1}{\delta}$ are independent since $\chi^2(Q\|P)$ is not multiplied by a factor of $\frac{1}{\delta}$.

C.2 Sub-Gaussian Loss Function

In some contexts, such as regression, considering bounded loss functions is restrictive. Next, we relax the restrictions on the loss function to deal with unbounded losses. We assume that, for any $h\in\mathcal{H}$, $\ell(h(x),y)$ is *sub-Gaussian.* First, we mention the definition of sub-Gaussian random variable Wainwright (2019).

Definition 2. *A random variable Z is said to be sub-Gaussian with the expectation $\mathbb{E}[Z]=\mu$ and variance proxy σ^2 if for any $\lambda\in\mathbb{R}$,*

$$\mathbb{E}[e^{\lambda(Z-\mu)}]\leq e^{\frac{\lambda^2\sigma^2}{2}}$$

Next, we present our PAC-Bayesian bounds.

Proposition 6 (The PAC-Bayesian bounds for sub-Gaussian loss function)**.** *Let P be a fixed prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size m and $\alpha > 1$, with probability at least $1 - \delta$, simultaneously for all posterior distributions Q, we have*

$$
\begin{aligned}
R_D(G_Q) &\leq R_S(G_Q) \\
&\quad + \sqrt{\frac{2\sigma^2}{m}\log(\frac{2}{\delta})\big(\alpha(\alpha-1)D_\alpha(Q\|P)+1\big)^{\frac{1}{\alpha}}}
\end{aligned}
$$

$$
\begin{aligned}
&R_D(G_Q) \leq R_S(G_Q) + \\
&\sqrt{\frac{1}{m}\bigg(D_\alpha(Q\|P)+\frac{1}{\alpha(\alpha-1)}\bigg)+\frac{1}{m\alpha}\bigg(2\sigma^2(\alpha-1)\log(\frac{2}{\delta})\bigg)^{\frac{\alpha}{\alpha-1}}}
\end{aligned}
$$

Proof. Suppose the convex function $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined as in Proposition 5. Employing Chernoff's bound, the tail bound probability for sub-Gaussian random variables Wainwright (2019) is given as follows

$$
Pr(|\bar{Z} - \mu| \geq \epsilon) \leq 2e^{-\frac{m\epsilon^2}{2\sigma^2}} \tag{17}
$$

Setting $\overline{Z} = R_S(h)$, $\mu = R_D(h)$ and $\delta = 2e^{-\frac{m\epsilon^2}{2\sigma^2}}$ in the tail bound in Equation (17), for any $h \in \mathcal{H}$, we have

$$
\begin{aligned}
&\underset{(x,y)\sim D}{Pr}(\phi_D(h) \geq \epsilon) \\
&= \underset{(x,y)\sim D}{Pr}\left((R_S(h) - R_D(h))^2 \geq \frac{\epsilon}{t}\right) \\
&= \underset{(x,y)\sim D}{Pr}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i), y_i) - \underset{(x,y)\sim D}{\mathbb{E}}\ell(h(x), y)\right| \geq \sqrt{\frac{\epsilon}{t}}\right) \\
&\leq 2e^{-\frac{m\epsilon}{2\sigma^2 t}}
\end{aligned}
$$

Setting $\delta = 2e^{-\frac{m\epsilon}{2\sigma^2 t}}$, we have

$$
\phi_D(h) \underset{1-\delta}{\leq} \frac{2t\sigma^2}{m}\log(\frac{2}{\delta})
$$

where $\phi_D(h)$ is defined as in Equation (13). By Equation (14), we have

$$
\underset{h\sim P}{\mathbb{E}}[(\phi_D(h))^{\frac{\alpha}{\alpha-1}}] \underset{1-\delta}{\leq} \left(\frac{2t\sigma^2}{m}\log(\frac{2}{\delta})\right)^{\frac{\alpha}{\alpha-1}} \tag{18}
$$

On the other hand, we may upper-bound $\mathbb{E}_{h\sim P}[(\phi_D(h))^{\frac{\alpha}{\alpha-1}}]$ in the following way (as in Proposition 6. in Alquier and Guedj (2018)).

$$
\begin{aligned}
\underset{h\sim P}{\mathbb{E}}[(\phi_D(h))^{\frac{\alpha}{\alpha-1}}] &\underset{1-\delta}{\leq} \frac{t^{\frac{\alpha}{\alpha-1}}}{\delta}\underset{(x,y)\sim D}{\mathbb{E}}\;\underset{h\sim P}{\mathbb{E}}[(R_S(h) - R_D(h))^{\frac{2\alpha}{\alpha-1}}] \\
&= \frac{t^{\frac{\alpha}{\alpha-1}}}{\delta}\underset{h\sim P}{\mathbb{E}}\;\underset{(x,y)\sim D}{\mathbb{E}}[(R_S(h) - R_D(h))^{\frac{2\alpha}{\alpha-1}}]
\end{aligned}
$$

Let $U = R_S(h) - R_D(h)$ and $q = \frac{\alpha}{\alpha-1}$. Then, U is a sub-Gaussian random variable from our assumption.

$$
\begin{aligned}
&\underset{(x,y)\sim D}{\mathbb{E}}[(R_S(h) - R_D(h))^{\frac{2\alpha}{\alpha-1}}] = \underset{(x,y)\sim D}{\mathbb{E}}[U^{2q}] \\
&= \int_0^\infty Pr\{|U|^{2q} > u\}du = 2q\int_0^\infty u^{2q-1}Pr\{|U| > u\}du \\
&\leq 4q\int_0^\infty u^{2q-1}e^{-\frac{mu^2}{2\sigma^2}}du
\end{aligned}
$$

By setting $u = \sqrt{t}$, the previous inequality becomes

$$\begin{aligned}\underset{(x,y)\sim D}{\mathbb{E}}[(R_S(h) - R_D(h))^{\frac{2\alpha}{\alpha-1}}] &\leq 2q \int_0^\infty t^{q-1} e^{-\frac{mt}{2\sigma^2}} dt \\ &= 2q\Gamma(q)\left(\frac{m}{2\sigma^2}\right)^{-q} = 2\frac{\alpha}{\alpha-1}\Gamma\left(\frac{\alpha}{\alpha-1}\right)\left(\frac{2\sigma^2}{m}\right)^{\frac{\alpha}{\alpha-1}}\end{aligned}$$

Therefore, we have

$$\underset{h\sim P}{\mathbb{E}}[(\phi_D(h))^{\frac{\alpha}{\alpha-1}}] \underset{1-\delta}{\leq} \frac{2\alpha t^{\frac{\alpha}{\alpha-1}}}{\delta(\alpha-1)}\Gamma\left(\frac{\alpha}{\alpha-1}\right)\left(\frac{2\sigma^2}{m}\right)^{\frac{\alpha}{\alpha-1}} \tag{19}$$

Although one has choices to use either Equation (18) or (19), we can easily see that Equation (18) is always tighter than (19). Putting everything together with Lemma 5, we prove our claim. □

The following corollary is an immediate consequnece of Proposition 6 for $\alpha = 2$.

Corollary 4 (The PAC-Bayesian bounds with χ^2-divergence for sub-Gaussian loss function)**.** *Let P be any prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size $m > 0$, with probability at least $1-\delta$, simultaneously for all posterior distributions Q, we have*

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{2\sigma^2}{m}\log(\frac{2}{\delta})\sqrt{\chi^2(Q\|P)+1}} \tag{20}$$

$$\begin{aligned}R_D(G_Q) &\leq R_S(G_Q) \\ &\quad + \sqrt{\frac{1}{2m}\left(\chi^2(Q\|P) + 1 + \left(2\sigma^2\log(\frac{2}{\delta})\right)^2\right)}\end{aligned} \tag{21}$$

Alquier and Guedj (2018) proved PAC-Bayes bound for sub-Gaussian loss function with χ^2-divergence. It is noteworthy that Δ may be any convex function and a different choice of Δ leads us to various bounds. For instance, choosing $\Delta(q,p) = |q-p|$ for Lemma 10 results in

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{2\sigma^2}{m}\log(\frac{2}{\delta})(\chi^2(Q\|P)+1)}$$

which is quite similar to the bounds in Proposition 6 in Alquier and Guedj (2018) and looser than Equation (20). We obtained a tighter bound due to our choice of $\Delta(q,p) = (q-p)^2$.

C.3 Sub-Exponential Loss Function

We now turn to a more general class where $\ell(h(x), y)$ is sub-exponential for any $h \in \mathcal{H}$. First, we define sub-exponentiality Wainwright (2019).

Definition 3. *A random variable Z is said to be sub-exponential with the expectation $\mathbb{E}[Z] = \mu$ and parameters σ^2 and $\beta > 0$, if for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda(Z-\mu)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}, \forall : |\lambda| < \frac{1}{\beta}$$

Next, we provide our PAC-Bayesian bounds.

Proposition 7 (The PAC-Bayesian bounds for sub-exponential loss function)**.** *Let P be a fixed prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size m and $\alpha > 1$, with probability at least $1-\delta$, simultaneously for all posterior distributions Q, we have*

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\mathcal{K}_\delta^1\big(\alpha(\alpha-1)D_\alpha(Q\|P)+1\big)^{\frac{1}{\alpha}}}$$

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1}{t}\bigg(D_\alpha(Q\|P) + \frac{1}{\alpha(\alpha-1)}\bigg) + \mathcal{K}^2_{\alpha,\delta}}$$

where

$$\mathcal{K}^1_\delta = \begin{cases} \frac{2\sigma^2}{m}\log(\frac{2}{\delta}), & \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2} \leq m \\ (\frac{2\beta}{m}\log\frac{2}{\delta})^2, & 0 < m < \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2}, \end{cases}$$

$$\mathcal{K}^2_{\alpha,\delta} = \frac{m^{\frac{1}{\alpha-1}}(\alpha-1)^{\frac{\alpha}{\alpha-1}}}{\alpha}(\mathcal{K}^1_\delta)^{\frac{\alpha}{\alpha-1}}$$

Proof. Suppose the convex function $\Delta : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is defined as in Proposition 5. For any random variables satisfying Definition 3, we have the following concentration inequality Wainwright (2019).

$$Pr(|\overline{Z}-\mu| \geq \epsilon) \leq \begin{cases} 2e^{-\frac{m\epsilon^2}{2\sigma^2}}, & 0 < \epsilon \leq \frac{\sigma^2}{\beta} \\ 2e^{-\frac{m\epsilon}{2\beta}}, & \frac{\sigma^2}{\beta} < \epsilon \end{cases} \tag{22}$$

Following the proof of Proposition 6, for $0 < \sqrt{\frac{\epsilon}{t}} \leq \frac{\sigma^2}{\beta}$, we have

$$\phi_D(h) \underset{1-\delta}{\leq} \frac{2t\sigma^2}{m}\log(\frac{2}{\delta})$$

For $\frac{\sigma^2}{\beta} < \sqrt{\frac{\epsilon}{t}}$, we have

$$\begin{aligned} &\underset{(x,y)\sim D}{Pr}(\phi_D(h) \geq \epsilon) \\ &= \underset{(x,y)\sim D}{Pr}\bigg((R_S(h)-R_D(h))^2 \geq \frac{\epsilon}{t}\bigg) \\ &= \underset{(x,y)\sim D}{Pr}\bigg(\bigg|\frac{1}{m}\sum_{i=1}^m \ell(h(x_i),y_i) - \underset{(x,y)\sim D}{\mathbb{E}}\ell(h(x),y)\bigg| \geq \sqrt{\frac{\epsilon}{t}}\bigg) \\ &\leq 2e^{-\frac{m}{2\beta}\sqrt{\frac{\epsilon}{t}}} \end{aligned}$$

Setting $\delta = 2e^{-\frac{m}{2\beta}\sqrt{\frac{\epsilon}{t}}}$, we have, for $\frac{\sigma^2}{\beta} < \sqrt{\frac{\epsilon}{t}}$,

$$\phi_D(h) \underset{1-\delta}{\leq} t\bigg(\frac{2\beta}{m}\log\frac{2}{\delta}\bigg)^2$$

Thus,

$$\phi_D(h) \underset{1-\delta}{\leq} \begin{cases} \frac{2t\sigma^2}{m}\log(\frac{2}{\delta}), & \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2} \leq m \\ t(\frac{2\beta}{m}\log\frac{2}{\delta})^2, & 0 < m < \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2} \end{cases}$$

where $\phi_D(h)$ is defined as in Equation (13). Now, we have the upper bound for $\phi_D(h)$ so we can apply the same procedure as in Proposition 5 and 7. □

Note that, for $\frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2} \leq m$, $\phi_D(h)$ behaves like sub-Gaussian. However, when the sample size is small, a tighter bound (i.e., $t(\frac{2\beta}{m}\log\frac{2}{\delta})^2$) can be obtained. This shows the advantage of assuming sub-exponentiality over sub-Gaussianity. By setting $\alpha = 2$ in Proposition 7, we have the following corollary.

Corollary 5 (The PAC-Bayesian bounds with χ^2-divergence for sub-exponential loss function)**.** *Let P be any prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size $m > 0$, with probability at least $1-\delta$, simultaneously for all posterior distributions Q, we have*

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\mathcal{K}^1_\delta\sqrt{\chi^2(Q\|P)+1}}$$

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1}{2m}\big(\chi^2(Q\|P) + 1 + (m\mathcal{K}_\delta^1)^2\big)}$$

where

$$\mathcal{K}_\delta^1 = \begin{cases} \frac{2\sigma^2}{m}\log(\frac{2}{\delta}), & \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2} \leq m \\ (\frac{2\beta}{m}\log\frac{2}{\delta})^2, & 0 < m < \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2} \end{cases}$$

To the best of our knowledge, our bounds for sub-exponential losses are entirely novel.

C.4 Proof of Proposition 1

Proof. Suppose the convex function $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined as in Proposition 5. By Chebyshev's inequality, for any $\epsilon > 0$ and any $h \in \mathcal{H}$,

$$\begin{aligned} &\underset{(x,y)\sim D}{Pr}(\phi_D(h) \geq \epsilon) \\ &= \underset{(x,y)\sim D}{Pr}\left((R_S(h) - R_D(h))^2 \geq \frac{\epsilon}{t}\right) \\ &= \underset{(x,y)\sim D}{Pr}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i), y_i) - \underset{(x,y)\sim D}{\mathbb{E}}\ell(h(x), y)\right| \geq \sqrt{\frac{\epsilon}{t}}\right) \\ &\leq \frac{t\sigma^2}{m\epsilon} \end{aligned}$$

Setting $\delta = \frac{t\sigma^2}{m\epsilon}$, we have

$$\phi_D(h) \underset{1-\delta}{\leq} \frac{t\sigma^2}{m\delta}$$

Now, we have the upper bound for $\phi_D(h)$ so we can apply the same procedure as in Proposition 5. □

D Log-concave distribution

We say that a distribution P with a density p (with respect to the Lebesgue measure) is a strongly log-concave distribution if the function $\log p$ is strongly concave. Equivalently stated, this condition means that the density can be expressed as $p(x) = \exp(-\psi(x))$, where the function $\psi : \mathbb{R}^d \to \mathbb{R}$ is strongly convex, meaning that there is some $\gamma > 0$ such that

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|x - y\|_2^2$$

for all $\lambda \in [0, 1]$, and $x, y \in \mathbb{R}^d$.