

## A Basic Differential Privacy

See (Dwork, 2006; Dwork and Roth, 2014) for more details.

Let  $U, U_1, U_2, \dots$  be independent uniform  $(0, 1)$  random variables, independent also of any other random variables unless stated otherwise, and let  $\pi : \mathbb{N} \times [0, 1] \rightarrow [0, 1]$  satisfy  $(\pi(1, U), \dots, \pi(k, U)) \stackrel{d}{=} (U_1, \dots, U_k)$  for all  $k \in \mathbb{N}$ . Write  $\pi_k$  for  $\pi(k, \cdot)$ .

**Definition A.1.** Let  $R, T$  be measurable spaces. A *randomized algorithm*  $\mathcal{A}$  from  $R$  to  $T$ , denoted  $\mathcal{A} : R \rightsquigarrow T$ , is a measurable map  $\mathcal{A} : [0, 1] \times R \rightarrow T$ . Associated to  $\mathcal{A}$  is a (measurable) collection of random variables  $\{\mathcal{A}_r : r \in R\}$  that satisfy  $\mathcal{A}_r = \mathcal{A}(U, r)$ . When there is no risk of confusion, we write  $\mathcal{A}(r)$  for  $\mathcal{A}_r$ .

For our purposes, we rely only on the fact that privacy is preserved under post-processing, which we now define.<sup>3</sup>

**Definition A.2.** Let  $\mathcal{A} : R \rightsquigarrow T$  and  $\mathcal{A}' : T \rightsquigarrow T'$ . The *composition*  $\mathcal{A}' \circ \mathcal{A} : R \rightsquigarrow T'$  is given by  $(\mathcal{A}' \circ \mathcal{A})(u, r) = \mathcal{A}'(\pi_2(u), \mathcal{A}(\pi_1(u), r))$ .

**Lemma A.3** (post-processing). *Let  $\mathcal{A} : Z^m \rightsquigarrow T$  be  $(\epsilon, \delta)$ -differentially private and let  $F : T \rightsquigarrow T'$  be arbitrary. Then  $F \circ \mathcal{A}$  is  $(\epsilon, \delta)$ -differentially private.*

## B Proof of Theorem 4.2

We prove a slightly more general result.

**Theorem B.1.** *Fix a bounded loss  $\ell \in [0, 1]$ . Let  $m \in \mathbb{N}$ , let  $\mathcal{P} : Z^m \rightsquigarrow \mathcal{M}_1(\mathbb{R}^p)$  be an  $\epsilon$ -differentially private data-dependent prior, let  $\mathcal{Q} \in \mathcal{M}_1(Z)$ , and let  $S \sim \mathcal{Q}^m$ . Then, for all  $\delta \in (0, 1)$  and  $\beta \in (0, \delta)$ , with probability at least  $1 - \delta$ ,*

$$\forall Q \in \mathcal{M}_1(\mathbb{R}^p), \text{kl}(\hat{L}_S(Q) \| L_{\mathcal{Q}}(Q)) \leq \frac{\text{KL}(Q \| \mathcal{P}(S)) + \ln \frac{2\sqrt{m}}{\delta - \beta}}{m} + \epsilon^2/2 + \epsilon \sqrt{\frac{\ln(2/\beta)}{2m}}. \quad (\text{B.1})$$

*Proof.* For every distribution  $P$  on  $\mathbb{R}^p$ , let

$$R(P) = \left\{ S \in Z^m : (\exists Q) \text{kl}(\hat{L}_S(Q) \| L_{\mathcal{Q}}(Q)) \geq m^{-1} (\text{KL}(Q \| P) + \ln \frac{2\sqrt{m}}{\delta'}) \right\}. \quad (\text{B.2})$$

It follows from Theorem 4.1 that  $\mathbb{P}_{S \sim \mathcal{Q}^m} \{S \in R(P)\} \leq \delta'$ . Let  $\beta > 0$ . Then, by the definition of approximate max-information, we have

$$\mathbb{P}_{S \sim \mathcal{Q}^m} \{S \in R(\mathcal{P}(S))\} \leq e^{I_{\infty}^{\beta}(\mathcal{P}; m)} \mathbb{P}_{(S, S') \sim \mathcal{Q}^{2m}} \{S \in R(\mathcal{P}(S'))\} + \beta \quad (\text{B.3})$$

$$\leq e^{I_{\infty}^{\beta}(\mathcal{P}; m)} \delta' + \beta \stackrel{\text{def}}{=} \delta. \quad (\text{B.4})$$

We have  $\delta' = e^{-I_{\infty}^{\beta}(\mathcal{P}; m)} (\delta - \beta)$ . Therefore, with probability no more than  $\delta$  over  $S \sim \mathcal{Q}^m$ ,

$$\exists Q \in \mathcal{M}_1(\mathbb{R}^p), \text{kl}(\hat{L}_S(Q) \| L_{\mathcal{Q}}(Q)) \geq \frac{\text{KL}(Q \| \mathcal{P}(S)) + \ln \frac{2\sqrt{m}}{\delta - \beta} + I_{\infty}^{\beta}(\mathcal{P}; m)}{m}. \quad (\text{B.5})$$

The result follows from replacing the approximate max-information  $I_{\infty}^{\beta}(\mathcal{P}; m)$  with the bound provided by Theorem 3.3.  $\square$

<sup>3</sup>It is sometimes more natural to refer to the differential privacy of probability kernels, i.e., measurable maps from  $Z^m$  to  $\mathcal{M}_1(T)$ , and to  $S$ -measurable random probability measures  $Q$ , i.e., probability kernels defined on the basic probability space satisfying  $Q = g(S)$  for some probability kernel  $g : Z^m \rightarrow \mathcal{M}_1(T)$ , where  $S \sim \mathcal{Q}^m$ . In both cases, the connection to the above definition is the same: for every probability kernel  $\kappa : R \rightarrow \mathcal{M}_1(T)$  there exists  $\mathcal{A} : [0, 1] \times R \rightarrow T$  such that  $\mathcal{A}(U, r) \sim \kappa(r)$  for every  $r \in R$ . In the other direction, clearly  $\kappa(r)(A) = \mathbb{P}\{\mathcal{A}(U, r) \in A\}$  for every measurable  $A \subseteq T$ .

The theorem leaves open the choice of  $\beta < \delta$ . For any fixed values for  $\epsilon$ ,  $m$ , and  $\delta$ , it is easy to optimize  $\beta$  to obtain the tightest possible bound. In practice, however, the optimal bound is almost indistinguishable from that obtained by taking  $\beta = \delta/2$ . For the remainder of the paper, we take this value for  $\beta$ , in which case, the r.h.s. of Eq. (4.2) is

$$\frac{\text{KL}(Q||\mathcal{P}(S)) + \ln \frac{4\sqrt{m}}{\delta}}{m} + \epsilon^2/2 + \epsilon \sqrt{\frac{\ln(4/\delta)}{2m}}. \quad (\text{B.6})$$

Note that the bound holds for all posteriors  $Q$ . In general the bounds are interesting only when  $Q$  is data dependent, otherwise one can obtain tighter bounds via concentration of measure results for empirical means of bounded i.i.d. random variables.

When one is choosing the privacy parameter,  $\epsilon$ , there is a balance between minimizing the direct contributions of  $\epsilon$  to the bound (forcing  $\epsilon$  smaller) and minimizing the indirect contribution of  $\epsilon$  through the KL term for posteriors  $Q$  that have low empirical risk (forcing  $\epsilon$  larger). One approach is to compute the value of  $\epsilon$  that achieves a certain bound on the excess generalization error. In particular, choosing  $\epsilon^2/2 = \alpha$  contributes an additional gap of  $\alpha$  to the KL-generalization error. Choosing  $\alpha$  is complicated by the fact that there is a non-linear relationship between the generalization error and the KL-generalization error, depending on the empirical risk. A better approach is often to attempt to balance the direct contribution with the indirect one. Regardless, the optimal value for  $\epsilon$  is much less than one, which can be challenging to obtain. We discuss strategies for achieving the required privacy in later sections.

## C Proofs for Section 5.1

These results connect approximations to differential privacy with bounds on the KL term in a PAC-Bayesian bound, yielding PAC-Bayes bounds that hold even if the prior is chosen via a nonprivate mechanism. In independent work, subsequent to our original arXiv preprint, Rivasplata et al. (2018) combined PAC-Bayesian bounds and stability to leverage *distribution* dependent priors. Their approach is distinct, though complimentary. See also work by London, 2017, who combines stability and PAC-Bayesian bounds in yet another way.

*Proof of Lemma 5.3.* Assume  $Q \ll P'$ , for otherwise the bound is trivial as  $\text{KL}(Q||P') = \infty$ . Then  $Q \ll P$ , because  $P' \ll P$ , and so  $\frac{dQ}{dP} = \frac{dQ}{dP'} \frac{dP'}{dP}$  and

$$\text{KL}(Q||P) - \text{KL}(Q||P') = Q \left[ \ln \frac{dQ}{dP} \right] - \text{KL}(Q||P') \quad (\text{C.1})$$

$$= Q \left[ \ln \frac{dQ}{dP'} + \ln \frac{dP'}{dP} \right] - \text{KL}(Q||P') \quad (\text{C.2})$$

$$= \text{KL}(Q||P') + Q \left[ \ln \frac{dP'}{dP} \right] - \text{KL}(Q||P') \quad (\text{C.3})$$

$$= Q \left[ \ln \frac{dP'}{dP} \right]. \quad (\text{C.4})$$

□

*Proof of Lemma 5.4.* Let  $P^*(S)$  satisfy the conditions in the statement of the theorem. Then  $P^*(S)$  is  $\epsilon$ -differentially private. By Theorem 4.2, the bound in Eq. (4.2) holds with probability at least  $1 - \delta$  for the data-dependent prior  $P^*(S)$  and all posteriors  $Q$ . By hypothesis, with probability  $1 - \delta - \delta'$ ,  $P^S \ll P^*(S)$ , and so, by Lemma 5.3,  $\text{KL}(Q||P^*(S)) = \text{KL}(Q||P^S) + Q \left[ \ln \frac{dP^S}{dP^*(S)} \right]$ . □

*Proof of Lemma 5.5.* Expanding the log ratio of Gaussian densities and then applying Cauchy–Schwarz, we obtain

$$\ln \frac{dN(w')}{dN(w)}(v) = \frac{1}{2} (\|w - v\|_{\Sigma^{-1}}^2 - \|w' - v\|_{\Sigma^{-1}}^2) \quad (\text{C.5})$$

$$= \langle w' - w, v \rangle_{\Sigma^{-1}} + \frac{1}{2} (\|w\|_{\Sigma^{-1}}^2 - \|w'\|_{\Sigma^{-1}}^2) \quad (\text{C.6})$$

$$= \frac{1}{2} \langle w' - w, 2v \rangle_{\Sigma^{-1}} - \frac{1}{2} \langle w' - w, w + w' \rangle_{\Sigma^{-1}} \quad (\text{C.7})$$

$$= \frac{1}{2} \langle w' - w, 2v - w - 2w' + w' \rangle_{\Sigma^{-1}} \quad (\text{C.8})$$

$$= \frac{1}{2} \langle w' - w, 2(v - w') + w' - w \rangle_{\Sigma^{-1}} \quad (\text{C.9})$$

$$\leq \frac{1}{2} \|w' - w\|_{\Sigma^{-1}}^2 + \|w' - w\|_{\Sigma^{-1}} \|v - w'\|_{\Sigma^{-1}}. \quad (\text{C.10})$$

The result follows by taking the expectation with respect to  $v \sim Q$ .  $\square$

*Proof of Lemma 5.6.* Let  $g = \frac{dQ}{dP} \stackrel{\text{def}}{=} \frac{e^h}{P[e^h]}$ . Then  $\|g\|_{L^1(P)} = 1$  and  $\|g\|_{L^\infty(P)} \leq e^{\|h\|_{L^\infty(P)}}$ . Let  $f(v) = \|v - w\|_{\Sigma^{-1}}$ . Then  $\mathbb{E}_{v \sim Q} \|v - w\|_{\Sigma^{-1}} = \|f\|_{L^1(Q)} = \|fg\|_{L^1(P)}$ . Finally, let  $\chi$  be the indicator function for the ellipsoid  $\{v : \|v - w\|_{\Sigma^{-1}} \leq R\}$ , and let  $\tilde{\chi} = 1 - \chi$ . Then  $\|f\chi\|_{L^\infty(P)} \leq R$  and

$$\|fg\|_{L^1(P)} = \|fg\chi\|_{L^1(P)} + \|fg\tilde{\chi}\|_{L^1(P)} \quad (\text{C.11})$$

$$\leq \|f\chi\|_{L^\infty(P)} \|g\|_{L^1(P)} + \|f\tilde{\chi}\|_{L^1(P)} \|g\|_{L^\infty(P)} = R + \sqrt{\frac{2}{\pi}} e^{-\frac{R^2}{2}} e^{\|h\|_{L^\infty(P)}}, \quad (\text{C.12})$$

where the inequalities follow from two applications of Hölder's inequality. Choosing  $R = \sqrt{2\|h\|_{L^\infty(P)}}$  gives  $\|f\|_{L^1(Q)} \leq \sqrt{2\|h\|_{L^\infty(P)}} + \sqrt{2/\pi}$ .  $\square$

*Proof of Corollary 5.7.* Let  $P^S = N(w(S))$  and  $P^*(S) = N(w^*(S))$ . By the closure of  $\varepsilon$ -differential privacy under composition,  $P^*(S)$  is  $\varepsilon$ -differentially private and is absolutely continuous with respect to  $N(w)$  for all  $w$ , and so satisfies the conditions of Lemma 5.4. In particular, with probability  $1 - \delta$ , Eq. (4.2) holds with  $\text{KL}(Q||P^*(S))$  replaced by  $\text{KL}(Q||P^S) + Q[\ln \frac{dP^S}{dP^*(S)}]$ .

By hypothesis, with probability at least  $1 - \delta - \delta'$ , it also holds that  $\|w(S) - w^*(S)\|_2^2 \leq C$ . Then, by Lemma 5.5,

$$Q \left[ \ln \frac{dP^S}{dP^*(S)} \right] \leq \frac{1}{2} \|w(S) - w^*(S)\|_2^2 / \sigma_{\min} + \|w(S) - w^*(S)\|_2 / \sqrt{\sigma_{\min}} \mathbb{E}_{v \sim Q} \|v - w(S)\|_{\Sigma^{-1}} \quad (\text{C.13})$$

$$\leq \frac{1}{2} C / \sigma_{\min} + \sqrt{C / \sigma_{\min}} \mathbb{E}_{v \sim Q} \|v - w(S)\|_{\Sigma^{-1}}. \quad (\text{C.14})$$

By Lemma 5.6  $\mathbb{E}_{v \sim Q} \|v - w(S)\|_{\Sigma^{-1}}$  is bounded by Gibbs measures based on a surrogate risk taking values in a length- $\Delta$  interval by  $\sqrt{2\tau\Delta} + \sqrt{2/\pi}$ .  $\square$

## D Bounded cross entropy

In order to achieve differential privacy, we work with a bounded version of the cross entropy loss. The problem is associated with extreme probabilities near zero and one. Our solution is to remap the probabilities  $p \mapsto \psi(p)$ , where

$$\psi(p) = e^{-\ell_{\max}} + (1 - 2e^{-\ell_{\max}})p \quad (\text{D.1})$$

is an affine transformation that maps  $[0, 1]$  to  $[e^{-\ell_{\max}}, 1 - e^{-\ell_{\max}}]$ , removing extreme probability values. Cross entropy loss is then replaced by  $g((p_1, \dots, p_K), y) = -\ln \psi(p_y)$ . As a result, cross entropy loss is contained in the interval  $[0, \ell_{\max}]$ . We take  $\ell_{\max} = 4$  in our experiments.

## E Computing PAC-Bayes bounds for Gibbs posteriors

For a given PAC-Bayes prior  $P$  and dataset  $S$ , it is natural to ask which posterior  $Q = Q(S)$  minimizes the PAC-Bayes bounds. In general, some Gibbs posterior (with respect to  $P$ ) is the minimizer. We now introduce the Gibbs posterior and discuss how we can compute the term  $\text{KL}(Q||P)$  in the case of Gibbs posteriors.

For a  $\sigma$ -finite measure  $P$  over  $\mathbb{R}^p$  and function  $g : \mathbb{R}^p \rightarrow \mathbb{R}$ , let  $P[g]$  denote the expectation  $\int g(h)P(dh)$  and, provided  $P[g] < \infty$ , let  $P_g$  denote the probability measure on  $\mathbb{R}^p$ , absolutely continuous with respect to  $P$ , with Radon–Nikodym derivative  $\frac{dP_g}{dP}(h) = \frac{g(h)}{P[g]}$ . A distribution of the form  $P_{\exp(-\tau g)}$  is generally referred to as a Gibbs distribution. A Gibbs *posterior* is a probability measure of the form  $P_{\exp(-\tau \hat{L}_S)}$  for some constant  $\tau > 0$ .

The challenge of evaluating PAC-Bayes bounds for Gibbs posteriors is computing the KL term. We now describe a classical estimate and show that it is going to be an upper bound with high probability. Fix a prior  $P$  and  $\tau \geq 0$ , let  $Q_\tau = P_{\exp(-\tau \hat{L}_S)}$ , and let  $Z_\tau = P[\exp(-\tau \hat{L}_S)]$ . Then

$$\text{KL}(Q_\tau||P) = Q_\tau \left[ \ln \frac{dQ_\tau}{dP} \right] \quad (\text{E.1})$$

$$= Q_\tau \left[ \ln \frac{\exp(-\tau \hat{L}_S)}{Z_\tau} \right] \quad (\text{E.2})$$

$$= -\tau Q_\tau[\hat{L}_S] - \ln Z_\tau. \quad (\text{E.3})$$

Letting  $W_1, \dots, W_n \sim Q_\tau$ , we have

$$Q_\tau[\hat{L}_S] = \sum_{i=1}^n Q_\tau[\hat{L}_S] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \hat{L}_S(W_i) \right]. \quad (\text{E.4})$$

(The quantity within the expectation on the r.h.s. thus defines an unbiased estimator of  $Q_\tau$ .) In the ideal case, the samples are independent, and then the variance decays at an  $n^{-1}$  rate. In practice, it is often difficult to even sample from  $Q_\tau$  for high values of  $\tau$ . Indeed, using this approach, we would generally overestimate the risk, which means that we do not obtain an upper bound on the KL term. So instead, we approximate  $-\tau Q_\tau[\hat{L}_S] \approx 0$ . Despite this, we obtain nonvacuous bounds. (For an alternative approach to this problem, see (Thiemann et al., 2017).)

The second term is challenging to estimate accurately, even assuming that  $P$  and  $Q_\tau$  can be efficiently simulated. One tack is to consider i.i.d. samples  $V_1, \dots, V_n \sim P$ , and note that

$$-\ln Z_\tau = -\ln P[\exp(-\tau \hat{L}_S)] = -\ln \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \exp(-\tau \hat{L}_S(V_i)) \right] \quad (\text{E.5})$$

$$\leq \mathbb{E} \left[ -\ln \frac{1}{n} \sum_{i=1}^n \exp(-\tau \hat{L}_S(V_i)) \right], \quad (\text{E.6})$$

where the inequality follows from an application of Jensen’s inequality. The quantity within the expectation on the r.h.s. thus forms an upper bound, and indeed, it is possible to show that it does not fall below the l.h.s. by  $\varepsilon$  with probability exponentially small in  $\varepsilon$ . Thus we have a high-probability (near) upper bound on the term in the KL. One might be inclined to compute a normalized importance sampler, but since  $Q$  cannot be effectively sampled, one does not obtain an upper bound with high probability.

The term  $\ln Z_\tau$  is a generalized log marginal likelihood, which, in our experiments, we approximate by sampling from a Gaussian distribution  $P$ . Numerical integration techniques rapidly diminish in accuracy with increasing dimensionality of the parameter space.

Note, that due to the convexity of the exponential, samples  $W_i \sim P$ , for which  $\hat{L}_S(W_i)$  is close to zero, will dominate  $Z_\tau$ . Due to high dimensionality of the neural network parameter space, with high probability a random sample  $W_i$  from  $P$  will not be far from minima of the empirical loss surface and therefore  $\hat{L}_S(W_i)$  will be high. As a results, in our experiments we obtain a very loose upper bound on the KL.

## F Experimental setup

**Bounded loss** While it is typical to train neural networks by minimizing cross entropy, this loss is unbounded and our theory is developed only for bounded loss. We therefore work with a bounded version of cross-entropy loss, which we obtain by preventing the network from producing extreme probabilities near zero and one. We describe our modification of the cross entropy in Appendix D.

**Datasets** We use two datasets. The first is MNIST, which consists of handwritten digit images with labels in  $\{0, \dots, 9\}$ . The dataset contains 50,000 training images and 10,000 validation images.

We also use a small synthetically generated dataset, which we refer to as SYNTH. The SYNTH dataset consists of 50 training data and 100 heldout data. Each input is a 4-dimensional vector sampled independently from a zero-mean Gaussian distribution with an identity covariance matrix. The true classifier is linear. The norm of the separating hyperplane is sampled from a standard normal.

The random label experiments are performed on a dataset where the labels are independently and uniformly generated and thus the risk is 0.5 under 0–1 loss.

**Architectures** We use SGLD without any standard modifications (such as momentum and batch norm) to ensure that the stationary distribution is that of SGLD. For MNIST, we use a fully connected neural network architecture. The network has 3 layers and 600 units in each hidden layer. The input is a 784 dimensional vector and the output layer has 10 units. For the SYNTH dataset, we use a fully connected neural network with 1 hidden layer consisting of 100 units. The input layer has 4 units, and the output layer is a single unit.

**Learning rate** At epoch  $t$ , the learning rate is  $a_t = a_0 * t^{-b}$ , where  $a_0$  is the initial learning rate and  $b$  is the decay rate. We set  $b = 0.5$  and use  $a_0 = 10^{-5}$  for MNIST experiments and  $a_0 = 10^{-3}$  for SYNTH experiments.

**Minibatches** An epoch refers to the full pass through the data in mini batches of size 128 for MNIST data, and 10 for SYNTH data.