7 Privacy-Preserving Mechanisms for SVM Learning

High-profile privacy breaches have trained the spotlight of public attention on data privacy. Until recently privacy, a relative laggard within computer security, could be enhanced only weakly by available technologies, when releasing aggregate statistics on sensitive data: until the mid-2000s definitions of privacy were merely syntactic. Contrast this with the state of affairs within cryptography that has long offered provably strong guarantees on maintaining the secrecy of encrypted information, based on computational limitations of attackers. Proposed as an answer to the challenge of bringing privacy on equal footing with cryptography, differential privacy (Dwork et al. 2006; Dwork & Roth 2014) has quickly grown in stature due to its formal nature and guarantees against powerful attackers. This chapter continues the discussion begun in Section 3.7, including a case study on the release of trained support vector machine (SVM) classifiers while preserving training data privacy. This chapter builds on (Rubinstein, Bartlett, Huang, & Taft 2012).

7.1 Privacy Breach Case Studies

We first review several high-profile privacy breaches achieved by privacy researchers. Together these have helped shape the discourse on privacy and in particular have led to important advancements in privacy-enhancing technologies. This section concludes with a discussion of lessons learned.

7.1.1 Massachusetts State Employees Health Records

An early privacy breach demonstrated the difficulty in defining the concept of *personally identifiable information* (PII) and led to the highly influential development of k-anonymity (Sweeney 2002).

In the mid-1990s the Massachusetts Group Insurance Commission released private health records of state employees, showing individual hospital visits, for the purpose of fostering health research. To mitigate privacy risks to state employees, the Commission scrubbed all suspected PII: names, addresses, and Social Security numbers. What was released was pure medical information together with (what seemed to be innocuous) demographics: birthdate, gender, and zipcode.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Security researcher Sweeney realized that the demographic information not scrubbed was in fact partial PII. To demonstrate her idea, Sweeney obtained readily available public voter information for the city of Cambridge, Massachusetts, which included birthdates, zipcodes, and names. She then linked this public data to the "anonymized" released hospital records, thereby re-identifying many of the employees including her target, then Governor William Weld who originally oversaw the release of the health data.

Sweeney (2002) measured the success rate of her technique more broadly, estimating that 87% of the U.S. population was uniquely identified by birthdate, gender, and zipcode. Focusing on preventing such unique identification, she proposed k-anonymity: the information on each individual in a release must be indistinguishable to at least $k-1$ other individuals in the release. Typically attributes/values identified as PII are either suppressed completely, or aggregated in a higher level of quantization.

Since the initial work on k-anonymity, vulnerabilities in the definition have been identified and in some cases repaired by follow-up derivative proposals such as ℓ-diversity (Machanavajjhala et al. 2007) and t-closeness (Li, Li, & Venkatasubramanian 2007).

7.1.2 AOL Search Query Logs

In 2006 AOL publicly released three months of search logs for over 650,000 AOL users, for the purpose of fostering Web search research (Barbaro & Zeller 2006). While queries were not labeled by user name, queries were linked to common user IDs, and search terms included PII and sensitive information regarding users' online activity.

Shortly after the data release, *New York Times* journalists identified AOL users in the query log. The misstep of releasing the data cost AOL's CTO her job and resulted in a class-action lawsuit brought against AOL by the affected users. The AOL search data episode put a dampener on data sharing for research, and highlights the difficulty in defining PII only in terms of structured data fields like name and address.

7.1.3 The Netflix Prize

In 2006—the same year as the AOL data release—Netflix launched a three-year-long movie recommendation competition with a $1,000,000 first prize (Bennett, Lanning et al. 2007). The released competition data consisted of about 100 million ratings of about 17 thousand movies by 480k users. The Netflix dataset was highly sparse with over 200 ratings on average per user and over 5k ratings on average per movie, but high variance on both counts.

While Netflix did not release information directly identifying users—indeed the ratings of some users were apparently perturbed in an attempt to preserve privacy (Bennett et al. 2007)—the identities of movies were released. Competitors were permitted to leverage movie identities to make use of external data that could assist in making more accurate recommendations. Researchers Narayanan & Shmatikov (2008) used publicly available IMDb data with movie ratings from non-anonymized users, to

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

link PII in public IMDb profiles with anonymized users in the Netflix data, thereby re-identifying users in the competition dataset. The sexual orientation of one Netflix user became public. Netflix lost a class-action lawsuit launched as a result of the de-anonymization, with one stipulation being the cancellation of the sequel Netflix Prize II competition.

7.1.4 Deanonymizing Twitter Pseudonyms

A year after Narayanan & Shmatikov re-identified Netflix users, the same researchers demonstrated the inherent difficulty of anonymizing social network data (Narayanan & Shmatikov 2009).

Privacy is a serious concern in social networks. For example, Twitter played a crucial role in the Arab Spring uprising, with political dissidents taking to the online site to coordinate peaceful protest and document government abuses. Pseudonyms in online profiles promote anonymity, but are not foolproof. And while companies like Twitter may attempt to keep user privacy protected, they have competing drives for profit (sharing data with advertisers) and complying with legal warrants for information. Narayanan & Shmatikov showed that even in the absence of any profile information, network connections alone can identify many users. Linking nodes in the Twitter social graph with nodes in the Flickr photo-sharing website's social graph, they were able to re-identify a third of the users verified to have accounts on both Twitter and Flickr with 88% accuracy.

We have extended the approach, with Narayanan and Shi, to recover test labels in a social network link prediction challenge (Narayanan, Shi, & Rubinstein 2011): effectively a privacy attack applied to cheating at a machine learning competition. Where Narayanan & Smatikov linked different graphs from two sites crawled at the same time, our result demonstrates that even over a six-month period graph evolution is insufficient to mitigate linkage attacks.

7.1.5 Genome-Wide Association Studies

Genome-wide Association Studies (GWAS) are an important statistical tool for analyzing medical assay data, in which the nucleotide frequencies at specific locii on the genome are compared between a control group of healthy subjects and a case group of individuals with a particular disease. Significant differences in occurrence frequencies imply an association between the disease and genetic markers. Homer et al. (2008) demonstrated the possibility of detecting whether a target individual participated in a GWAS case group—characterized as a mixture of participant genetic data—from the target DNA sequence information. This opened the possibility that participants of previously published GWAS data could be identified and led the National Institutes of Health to control the publication of study statistics.

Several other avenues exist for breaching "genetic privacy" (Erlich & Narayanan 2014) such as triangulating genealogical data to determine a study participant's surname (Gymrek, McGuire, Golan, Halperin, & Erlich 2013).

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

7.1.6 Ad Microtargeting

Sharing personal data such as demographics and browsing history with online advertisers has long led to objections from privacy advocates. But what if no directly identifying information is shared with advertisers? What if no information is directly shared at all? Even in this seemingly innocuous setting, user privacy may be breached. Korolova (2011) showed that ad microtargeting on Facebook can be exploited to infer private information on users. The basic attack aims to infer some unknown private attribute (such as sexual orientation) of a selected target, for whom some identifying information is known already, such as age, location, workplace, or education. The attacker launches two ad campaigns on the social network, both targeting the fine details of the target user: one targeting users having the private attribute and the other targeting users without the attribute. Even though Facebook does not *directly* share the details of users viewing the ad, the attacker can determine this information depending on which ad campaign's dollar balance changes. By examining the PII that is provided by Facebook, both campaigns can target the specific user, and an ad from exactly one of the two campaigns will be presented and be charged back to the attacker. This reveals which attribute value is possessed by the target user.

7.1.7 Lessons Learned

A number of patterns recur within the above case studies:

- It is dangerous to dismiss privacy breaches as trivially unlikely or requiring too much work on the part of the attacker. Time and time again, highly sophisticated attacks have been demonstrated. They were publicized to highlight misuses of sensitive data (due to initial data releases) and advance privacy research. In reality, much more valuable incentives exist to breach privacy without disclosure.
- Many breaches are accomplished via linkage attack: Records of an "anonymized" dataset are linked with an external data source that is easily obtained or even public. The combined data ties sensitive information (from the released private data) to re-identified individuals (from public records).
- It is exceedingly difficult to identify attributes that constitute PII, a necessary condition of employing *syntactic* measures such as k-anonymity and its derivatives. Many attributes can act like fingerprints: a *curse of dimensionality* exists in private data analysis in which high-dimensional data associated with individuals is likely to be unique.

7.2 Problem Setting: Privacy-Preserving Learning

Our goal in this chapter is to release aggregate information about a dataset while maintaining the privacy of each individual datum. These two goals of *utility* and *privacy* are inherently discordant. However, we will see that an effective balance can be achieved with practical release mechanisms.

For a release mechanism to be useful, its responses must closely resemble some target statistic of the data. In particular since we will be releasing classifiers learned on

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

training data, our formal measure of utility will compare the released classifiers with a desired nonprivate classifier trained on the same data. In this case, our target classifier is the support vector machine (SVM), one of the most widely used supervised learners in practice.

Within the context of the previous section, any claim of data confidentiality *must* be backed up by a semantic guarantee of privacy: Time and time again weakly anonymized data has been re-identified. To this end, it is necessary for the mechanism's response to be "smoothed out": the mechanism must be randomized to reduce any individual datum's influence on this distribution. This is exactly the approach when establishing a property due to Dwork et al. (2006) known as differential privacy. We adopt this strong guarantee on privacy.

Those in the area of statistical databases, studied by the databases and theory communities, hope to understand when the goals of utility and privacy can be efficiently achieved simultaneously (Dinur & Nissim 2003; Barak et al. 2007; Blum et al. 2008; Chaudhuri & Monteleoni 2009; Kasiviswanathan et al. 2008; Duchi et al. 2013; Dwork & Roth 2014; Aldà & Rubinstein 2017). We will thus adopt their terminology while examining their approaches. While this chapter's discussion involves theoretical analyses, the mechanisms presented here—developed in our past work (Rubinstein, Bartlett, Huang, & Taft 2009; Rubinstein et al. 2012)—are easily implemented and efficient, and release classifiers that are close to the corresponding nonprivate SVM under the ℓ_∞-norm, with high probability. In our setting this notion of utility is stronger than closeness of risk.

7.2.1 Differential Privacy

We now cover background on differential privacy. Given access to database $\mathbb{D}$, a *mechanism* M must release aggregate information about $\mathbb{D}$ while maintaining the privacy of individual entries. We assume that the *response* $M(\mathbb{D})$, belonging to range space $\mathcal{T}_M$, is the only information released by the mechanism. The statistical databases terminology we adopt should be understood to have analogs in machine learning, where for example database corresponds to dataset, record or item corresponds to datum, mechanism to learning algorithm, and the released quantity as the learned classifier.

We say that a pair of databases $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$ are neighbors if they differ on exactly one entry; in other words the two datasets are separated by hamming (or ℓ_1) distance 1. We adopt the following strong notion of privacy due to Dwork et al. (2006).

DEFINITION 7.1 For any $\beta > 0$, a randomized mechanism M provides *β-differential privacy*, if, for all neighboring databases $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$ and all responses $t \in \mathcal{T}_M$, the mechanism satisfies

$$\log\left(\frac{\Pr\left(M\left(\mathbb{D}^{(1)}\right)=t\right)}{\Pr\left(M\left(\mathbb{D}^{(2)}\right)=t\right)}\right) \leq \beta.$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

The probability in the definition is over the randomization in M, not the databases. For continuous $\mathcal{T}_M$ we mean by this ratio a Radon-Nikodym derivative of the distribution of $M\left(\mathbb{D}^{(1)}\right)$ with respect to the distribution of $M\left(\mathbb{D}^{(2)}\right)$.[1] In the sequel we assume without loss of generality that each pair of neighboring databases differs on their last entry.

To understand the definition, consider a mechanism M preserving a high level of differential privacy (low β) and a powerful attacker with the following capabilities:

- Knowledge of the mapping M up to randomness;
- Unbounded computational resources that could be used to simulate M's responses on any number of databases;
- Knowledge of the first $N-1$ entries of $\mathbb{D}$; and
- The ability to sample linearly many responses from $M(\mathbb{D})$.

Then the attacker's optimal approach to re-identify the N^{th} entry of $\mathbb{D}$ is to

1 Construct an empirical distribution approximating the unknown response distribution $M(\mathbb{D})$ by querying the mechanism the maximum order of N times and forming a histogram of responses;
2 Construct all candidate databases $\mathbb{D}'$ from the known $N-1$ first entries of $\mathbb{D}$;
3 For each $\mathbb{D}'$:
   1. Simulate the response distribution of $M(\mathbb{D}')$ using knowledge of $M(\cdot)$;
   2. Compare the exact candidate response distribution with the approximate empirical distribution; and
4 Return the $\mathbb{D}'$ that most closely matches the empirical distribution.

Even with this optimal procedure the adversary cannot infer additional information on the true identity of the N^{th} entry of $\mathbb{D}$, under suitably small β. Differential privacy guarantees that each response distribution $M(\mathbb{D}')$ is pointwise close to the true distribution $M(\mathbb{D})$. In other words the response distributions are indistinguishable because the unknown datum is varied over all possibilities. When attempting to match candidate response distribution to the queried actual empirical distribution, it will not be possible to distinguish true $\mathbb{D}$ from incorrect neighboring $\mathbb{D}'$. Indistinguishability is effective, provided β is close to the estimation error between the sampled empirical distribution and the actual response distribution $M(\mathbb{D})$—this estimation error is guaranteed not to be too small since the number of queries is limited.

We have argued in this section that differential privacy provides a strong, semantic guarantee of privacy for each individual datum, while allowing the release of some aggregate information. Indeed in certain situations one may argue the definition is unnecessarily strong. Some possible relaxations are natural; for example a "grouped privacy" variation models an adversary with knowledge of the data up to some fixed $k \geq 1$ rows of $\mathbb{D}$ (Dwork et al. 2006). In other words knowledge of the data is relaxed to a hamming-k ball centered on $\mathbb{D}$.

[1] More generally, for each measurable $T \subseteq \mathcal{T}_M$, $\Pr\left(M\left(\mathbb{D}^{(1)}\right) \in T\right) \leq \exp(\beta)\Pr\left(M\left(\mathbb{D}^{(2)}\right) \in T\right)$.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

7.2.1.1 The Laplace Mechanism

The earliest pattern of establishing differential privacy is to add (possibly multivariate) zero-mean Laplace noise to a nonprivate mechanism. Typically the nonprivate mechanism being privatized is a deterministic function of data—in our case it will be the support vector machine. The scale of the zero-mean Laplace noise depends on the level β of differential privacy desired and the so-called global sensitivity of the nonprivate mechanism.

DEFINITION 7.2 Deterministic mechanism M with real-vector-valued responses, has ℓ_1-global sensitivity $\Delta > 0$ if for every pair $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$ of neighboring databases, $\left\| M\left(\mathbb{D}^{(1)}\right) - M\left(\mathbb{D}^{(2)}\right) \right\|_1 \leq \Delta$.

This is essentially a Lipschitz condition on the nonprivate mechanism with metric in the co-domain induced by the ℓ_1-norm, and similarly on databases in the domain. Intuitively global sensitivity measures the continuity of the nonprivate map: how much does the response to be released vary with small perturbations to the input dataset.

With global sensitivity of a nonprivate mechanism in hand, it is a simple matter to prove differential privacy of suitable Laplace noise added to the nonprivate mechanism.

LEMMA 7.3 (Dwork et al. 2006) *Let M be a deterministic mechanism with ℓ_1-global sensitivity $\Delta > 0$, let $\beta > 0$, and $\lambda \overset{iid}{\sim} Laplace(\mathbf{0}, \beta/\Delta)$. Then mechanism $M(\mathbb{D}) + \lambda$ preserves β-differential privacy.*

Proof Consider neighboring databases $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$, any $t \in \mathcal{T}_M$, and two multivariate random variables $\lambda_1, \lambda_2 \overset{iid}{\sim} Laplace(\mathbf{0}, \Delta/\beta)$

$$\begin{aligned}\frac{\Pr\left(M\left(\mathbb{D}^{(1)}\right) + \lambda_1 = t\right)}{\Pr\left(M\left(\mathbb{D}^{(2)}\right) + \lambda_2 = t\right)} &= \frac{\exp\left(\left\| t - M\left(\mathbb{D}^{(1)}\right)\right\|_1 / (\Delta/\beta)\right)}{\exp\left(\left\| t - M\left(\mathbb{D}^{(2)}\right)\right\|_1 / (\Delta/\beta)\right)} \\ &\leq \exp\left(\left\| M\left(\mathbb{D}^{(1)}\right) - M\left(\mathbb{D}^{(2)}\right)\right\|_1 / (\Delta/\beta)\right) \\ &\leq \exp(\beta).\end{aligned}$$

The equality follows from the definition of the Laplace probability density function, the first inequality follows from the triangle inequality, and the final inequality follows from the bound on global sensitivity. Taking logs of both sides yields β-differential privacy. □

7.2.2 Utility

Intuitively the more an "interesting" mechanism M is perturbed to guarantee differential privacy, the less like M the resulting mechanism $\hat{M}$ will become. The next definition formalizes the notion of "likeness."

DEFINITION 7.4 Consider two mechanisms $\hat{M}$ and M with the same domains and with response spaces $\mathcal{T}_{\hat{M}}$ and $\mathcal{T}_M$. Let $\mathcal{X}$ be some set and let $\mathcal{F} \subseteq \Re^{\mathcal{X}}$ be parametrized by the response spaces: For every $t \in \mathcal{T}_{\hat{M}} \cup \mathcal{T}_M$ define some corresponding function $f_t \in \mathcal{F}$. Finally assume $\mathcal{F}$ is endowed with norm $\| \cdot \|_{\mathcal{F}}$. Then for $\epsilon > 0$ and $0 < \delta < 1$ we say

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

that[2] $\hat{M}$ is (ϵ, δ)-*useful* with respect to M if, for all databases $\mathbb{D}$,

$$\Pr\left(\left\| f_{\hat{M}(\mathbb{D})} - f_{M(\mathbb{D})} \right\|_{\mathcal{F}} \leq \epsilon\right) \geq 1 - \delta.$$

Typically $\hat{M}$ will be a privacy-preserving (perturbed) version of M, such as the Laplace mechanism of M. In the sequel we take $\|\cdot\|_{\mathcal{F}}$ to be $\|f\|_{\infty;\mathcal{M}} = \sup_{\mathbf{x}\in\mathcal{M}} |f(\mathbf{x})|$ for some $\mathcal{M} \subseteq \Re^D$ containing the data. It will also be convenient to define $\|k\|_{\infty;\mathcal{M}} = \sup_{\mathbf{x},\mathbf{z}\in\mathcal{M}} |k(\mathbf{x}, \mathbf{z})|$ for bivariate $k(\cdot, \cdot)$.

7.2.3 Historical Research Directions in Differential Privacy

A rich literature of prior work on differential privacy exists. We provide an overview some of this work in the context of privacy-preserving learning of support vector machines.

Range Spaces Parametrizing Vector-Valued Statistics or Simple Functions
Early work on private interactive mechanisms focused on approximating real- and vector-valued statistics (e.g., Dinur & Nissim 2003; Blum et al. 2005; Dwork et al. 2006; Dwork 2006; Barak et al. 2007) McSherry & Talwar (2007) first considered private mechanisms with range spaces parametrizing sets more general than real-valued vectors and used such differentially private mappings for mechanism design. The first work to develop private mechanisms for releasing classifiers, which specifically regularized logistic regression, was due to Chaudhuri & Monteleoni (2009). There the mechanism's range space parametrizes the VC-dimension $D + 1$ class of linear hyperplanes in $\Re^D$. One of their mechanisms injects a random term into the primal objective to achieve differential privacy. Their simpler mechanism adds noise to the learned weight vector. Both of these approaches have analogs in SVM learning as explored in this chapter. The calculation of SVM sensitivity (see Section 7.4) presented here is a generalization of the derivation of the sensitivity of regularized logistic regression (Chaudhuri & Monteleoni 2009), to the setting of nondifferentiable loss functions, with the condition on the gradient replaced by the Lipschitz condition and the condition on the Hessian replaced by strong convexity. Kasiviswanathan et al. (2008) show that discretized concept classes can be PAC or agnostically learned privately, albeit via an inefficient mechanism. Blum et al. (2008) show that noninteractive mechanisms can privately release anonymized data such that utility is guaranteed over classes of predicate queries with polynomial VC-dimension, when the domain is discretized. Dwork et al. (2009) has since characterized when utility and privacy can be achieved by efficient noninteractive mechanisms. In this chapter we consider efficient mechanisms for private SVM learning, whose range spaces parametrize real-valued functions. One case considered here is learning with a RBF (or Gaussian) kernel, which corresponds to learning over a rich class of infinite dimension. Differential privacy has been explored under Bayesian probabilistic inference in the exact (Dimitrakakis, Nelson, Mitrokotsa,

[2] Our definition of (ϵ, δ)-usefulness for releasing a single function is analogous to the notion of the same name introduced by Blum et al. (2008) for anonymization mechanisms.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

& Rubinstein 2014; Zhang, Rubinstein, & Dimitrakakis 2016) and sampled settings (Wang et al. 2015). Other work develops generic mechanisms for function release under differential privacy: Hall, Rinaldo, & Wasserman (2013) add Gaussian process noise, yielding a weaker form of differential privacy, without general results on utility rates; Wang, Fan, Zhang, & Wang (2013) release privatized projections, in a trigonometric basis, of functions that are separable in the training data, similar to Zhang et al. (2012); Aldà & Rubinstein (2017) propose the Bernstein mechanism as a functional form of the Laplace mechanism for vector release, which achieves differential privacy and strong utility guarantees under general conditions of smoothness of the nonprivate function, by using iterated Bernstein polynomial approximation.

Practical Privacy-Preserving Learning (Mostly) via Subset Sums

Most work in differential privacy has focused on the deep analysis of mechanisms for relatively simple statistics (with histograms and contingency tables as explored by Blum et al. 2005 and Barak et al. 2007, respectively, as examples) and learning algorithms (e.g., interval queries and halfspaces as explored by Blum et al. 2008), or on constructing learning algorithms that can be decomposed into subset-sum operations (e.g., perceptron, k-NN, ID3 as described by Blum et al. 2005, and various recommender systems as shown by McSherry & Mironov 2009). Some early work in function release focuses on functions separable (expressable as sums) over training data (Zhang et al. 2012; Wang et al. 2013). By contrast, this chapter explores the more practical goal of SVM learning, which does not generally decompose into a subset sum (Rubinstein et al. 2012; Appendix A). It is also notable that the mechanisms here run in polynomial time.

The Privacy-Utilty Tradeoff

Like several early studies, we explore here the tradeoff between privacy and utility. Barak et al. (2007) present a mechanism for releasing contingency tables that guarantees differential privacy and also guarantees a notion of accuracy: with high probability all marginals from the released table are close in ℓ_1-norm to the true marginals. As mentioned earlier, Blum et al. (2008) develop a private noninteractive mechanism that releases anonymized data such that all predicate queries in a VC class take on similar values regardless of whether they are taken over the anonymized data and original data. Kasiviswanathan et al. (2008) consider utility as corresponding to PAC learning: with high probability the response and target concepts are close, averaged over the underlying measure.

Early negative results show that any mechanism providing overly accurate responses cannot be private (Dinur & Nissim 2003; Dwork & Yekhanin 2008; Beimel et al. 2010; Hardt & Talwar 2010). Dinur & Nissim (2003) show that if a noise of rate only $o\left(\sqrt{N}\right)$ is added to subset-sum queries on a database of bits, then an adversary can reconstruct a $1 - o(1)$ fraction of the bits. This threshold phenomenon says that if accuracy is too great, privacy cannot be guaranteed at all. A similar negative result can be shown for the case of private SVM learning: i.e., requiring very high accuracy with respect to the SVM prevents high levels of privacy. De (2012) summarizes approaches to lower-bounding utility under differential privacy such as the volumetric packing approach used here.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

The results presented here are qualitatively closer to those of Hardt & Talwar (2010) and Beimel et al. (2010). The former work finds almost matching upper and lower bounds for the tradeoff between differential privacy and accuracy through the lens of convex geometry, in the following setting that encompasses releasing histograms and recommender systems. Queries submitted to the interactive mechanism are linear mappings on a private database of reals. Nonprivate responses are the vector image of the query applied to the database, the mechanism's responses are a randomized version of this target image, and the mechanism's accuracy is the expected Euclidean distance between nonprivate also private and private responses. Beimel et al. (2010) focus on the notion of private learning (Kasiviswanathan et al. 2008) in which a private learner not only PAC learns but the release of its hypothesis is differentially private with respect to the training data. Beimel et al. (2010) delve into the sample complexity of private learning and demonstrate separation results between proper and improper private learning[3]—which do not exist for nonprivate PAC learning—and between efficient and inefficient proper private learners. Both papers consider negative results on the tradeoff between notions of utility and differential privacy. In SVM learning, concept classes are not necessarily linear or have finite VC-dimension.

The ϵ-packing proof technique used in the lower bound for SVM learning with the RBF kernel, although discovered independently, is similar to the technique used by Hardt & Talwar (2010) to establish lower bounds for their setting of privately responding to linear map queries. See also (De 2012) for a general account.

Connections between Stability and Differential Privacy

To prove differential privacy in this chapter, a proof technique is borrowed from algorithmic stability. In passing, Kasiviswanathan et al. (2008) predict a possible relationship between algorithmic stability and differential privacy; however, they do not describe in detail how to exploit this. Since then, Wang et al. (2016) have drawn connections between stability, learnability, and privacy.

7.3 Support Vector Machines: A Brief Primer

Empirical risk minimization (ERM) can lead to overfitting or poor generalization (risk of the minimizer), so in theory and practice it is more desirable to perform regularized empirical risk minimization, which minimizes the sum of the empirical risk and uses a regularization term that imposes a soft smoothness constraint on the classifier. A well-known example is the soft-margin support vector machine that has the following primal program, for convex loss $\ell(\cdot, \cdot)$,

$$\min_{\mathbf{w} \in \Re^F} \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{C}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right),$$

[3] A proper learner outputs a hypothesis from the target concept class.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

where for chosen *feature mapping* $\phi : \Re^D \to \Re^F$ taking points in input space $\Re^D$ to some (possibly infinite) F-dimensional *feature space*, and hyperplane normal $\mathbf{w} \in \Re^F$, we define

$$f_{\mathbf{w}}(\mathbf{x}) = \langle \phi(\mathbf{x}), \mathbf{w} \rangle.$$

Parameter $C > 0$ is the soft-margin parameter that controls the degree of regularization. Let $\mathbf{w}^\star$ denote an optimizing weight vector. Then predictions are taken as the sign of $f^\star(\mathbf{x}) = f_{\mathbf{w}^\star}(\mathbf{x})$.We will refer to both $f_{\mathbf{w}}(\cdot)$ and $\text{sign}(f_{\mathbf{w}}(\cdot))$ as *classifiers*, with the exact meaning apparent from the context.

An overview of the relevant details on learning follows (for full details, see, for example Burges 1998; Cristianini & Shawe-Taylor 2000; Schölkopf & Smola 2001; Bishop 2006) In order for the primal to be convex and the process of SVM learning to be tractable, $\ell(y, \hat{y})$ is chosen to be a loss function that is convex in $\hat{y}$. A common convex surrogate for the 0-1 loss and the loss most commonly associated with the SVM, is the hinge loss $\ell(y, \hat{y}) = \max[1 - y\hat{y}, 0]$ that upper bounds the 0-1 loss and is non-differentiable at $y\hat{y} = 1$. Other example losses include the square loss $(1 - y\hat{y})^2$ and the logistic loss $\log(1 + \exp(-y\hat{y}))$. We consider general convex losses in this chapter, and a detailed case study of private SVM learning under the hinge loss in Section 7.7.1.

Remark 7.5 We say that a learning algorithm is *universally consistent* if for all distributions μ it is *consistent*: its expected risk converges to the minimum achievable (Bayes) risk with increasing sample size (Devroye, Györfi, & Lugosi 1996). For universal consistency, the SVM's parameter C should increase like $\sqrt{N}$.

When F is large the solution may be more easily obtained via the dual. For example, the following is the dual formulation on the N dual variables for learning with hinge loss:

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \Re^N} \; & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) \\ \text{s.t.} \; & 0 \leq \alpha_i \leq \frac{C}{N} \; \forall i \in [n], \end{aligned} \tag{7.1}$$

where $k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ is the kernel function.

The vector of maximizing duals $\boldsymbol{\alpha}^\star$ parametrizes the function $f^\star = f_{\boldsymbol{\alpha}^\star}$ as

$$f_{\boldsymbol{\alpha}}(\cdot) = \sum_{i=1}^{N} \alpha_i y^{(i)} k\left(\cdot, \mathbf{x}^{(i)}\right).$$

The space of SVM classifiers endowed with the kernel function forms a reproducing kernel Hilbert space $\mathcal{H}$.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Algorithm 7.1 SVM

Inputs: database $\mathbb{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_{i=1}^{N}$ with $\mathbf{x}^{(i)} \in \Re^D$, $y^{(i)} \in \{-1, 1\}$; kernel $k : \Re^D \times \Re^D \to \Re$; convex loss ℓ; parameter $C > 0$.

1 $\boldsymbol{\alpha}^\star \leftarrow$ Solve the SVM's dual in Equation (7.1).
2 Return vector $\boldsymbol{\alpha}^\star$.

DEFINITION 7.6 A *reproducing kernel Hilbert space (RKHS)* is a Hilbert space[4] of real-valued functions on the space $\mathcal{X}$, which includes, for each point $\mathbf{x} \in \mathcal{X}$, a point-evaluation function $k(\cdot, \mathbf{x})$ having the reproducing kernel property $\langle f, k(\cdot, \mathbf{x})\rangle_{\mathcal{H}} = f(\mathbf{x})$ for all $f \in \mathcal{H}$.

In particular $\langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{z})\rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{z})$. The Representer Theorem (Kimeldorf & Wahba 1971) implies that the minimizer $f^\star = \operatorname{argmin}_{f\in\mathcal{H}} \left[\frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{C}{N}\sum_{i=1}^{n} \ell\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right)\right]$ lies in the span of the functions $k\left(\cdot, \mathbf{x}^{(i)}\right) \in \mathcal{H}$. Indeed the above dual expansion shows that the coordinates in this subspace are given by the $\alpha_i^\star y^{(i)}$. We define the SVM *mechanism* to be the dual optimization that responds with the vector $\boldsymbol{\alpha}^\star$, as described by Algorithm 7.1.

7.3.1 Translation-Invariant Kernels

A number of kernels/feature mappings have been proposed in the literature (Burges 1998; Cristianini & Shawe-Taylor 2000; Schölkopf & Smola 2001; Bishop 2006). The *translation-invariant kernels* are an important class of kernel that we study in the sequel (see Table 7.1 for examples).

DEFINITION 7.7 A kernel function of the form $k(\mathbf{x}, \mathbf{z}) = g(\mathbf{x} - \mathbf{z})$, for some function g, is called *translation invariant*.

Table 7.1 Example of translation-invariant kernels and their g functions as defined on the vector $\boldsymbol{\Delta} = \mathbf{x} - \mathbf{z}$

Kernel	$g(\boldsymbol{\Delta})$
RBF	$\exp\left(-\frac{\|\boldsymbol{\Delta}\|_2^2}{2\sigma^2}\right)$
Laplacian	$\exp\left(-\|\boldsymbol{\Delta}\|_1\right)$
Cauchy	$\prod_{i=1}^{D} \frac{2}{1+\Delta_i^2}$

[4] A Hilbert space is an inner-product space that is complete with respect to its norm-induced metric.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

7.3.2 Algorithmic Stability

In proving bounds on the differential privacy of our mechanisms for private SVM learning, we will exploit the uniform stability of regularized ERM as established by Bousquet & Elisseeff (2002).

Recall that we say that a pair of databases $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$ are *neighbors* if they differ on one entry, and we define the learning stability with respect to neighboring databases as follows.

DEFINITION 7.8 A *learning map* $\mathcal{A}$, that takes databases $\mathbb{D}$ to classifiers is said to have γ*-uniform stability* with respect to loss $\ell(\cdot, \cdot)$ if for all neighboring databases $\mathbb{D}, \mathbb{D}'$, the losses of the classifiers trained on $\mathbb{D}$ and $\mathbb{D}'$ are close on all test examples $\left\| \ell(\cdot, \mathcal{A}(\mathbb{D})) - \ell(\cdot, \mathcal{A}(\mathbb{D}')) \right\|_\infty \leq \gamma$.

Stability corresponds to smoothness of the learning map, and the concept is typically used in statistical learning theory to yield tight risk bounds, sometimes when class capacity-based approaches (such as VC-dimension-based approaches) do not apply (Devroye & Wagner 1979; Kearns & Ron 1999; Bousquet & Elisseeff 2002; Kutin & Niyogi 2002). Intuitively if a learning map is stable, then it is not overly influenced by noise and is less likely to suffer from overfitting.

7.4 Differential Privacy by Output Perturbation

We now focus on an output perturbation approach we developed with collaborators Bartlett, Huang, and Taft, based on the Laplace mechanism (Rubinstein, Bartlett, Huang, & Taft 2009; Rubinstein et al. 2012). We consider differentially private SVM learning with finite F-dimensional feature maps. We begin by describing the mechanism, then prove the range of noise parameters required to guarantee privacy (Theorem 7.10), and derive the conditions under which the mechanism yields close approximations to the nonprivate SVM (Theorem 7.11).

Algorithm 7.2 describes the PRIVATESVM-FINITE mechanism, which is an application of the Laplace mechanism (cf. Section 7.2.1.1): After forming the primal solution to the SVM—weight vector $\mathbf{w} \in \Re^F$—the mechanism adds i.i.d. zero-mean, scale λ, and Laplace noise to $\mathbf{w}$. Differential privacy follows from the ℓ_1-sensitivity Δ of $\mathbf{w}$ to data perturbations taking the Laplace-noise scale to be $\lambda = \Delta/\beta$.

To calculate sensitivity—the change in $\mathbf{w}$ with respect to the ℓ_1-norm when a training example is perturbed—we exploit the uniform stability of regularized ERM (cf. Definition 7.8).

LEMMA 7.9 *Consider a loss function $\ell(y, \hat{y})$ that is convex and L-Lipschitz in $\hat{y}$, and RKHS $\mathcal{H}$ induced by finite F-dimensional feature mapping ϕ with bounded kernel $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2$ for all $\mathbf{x} \in \Re^D$. For each database $\mathbb{S} = \left\{ \left(\mathbf{x}^{(i)}, y^{(i)} \right) \right\}_{i=1}^N$, define*

$$\mathbf{w}^{(\mathbb{S})} \in \operatorname*{argmin}_{\mathbf{w} \in \Re^F} \left[\frac{C}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, f_{\mathbf{w}}\left(\mathbf{x}^{(i)} \right) \right) + \frac{1}{2} \|\mathbf{w}\|_2^2 \right].$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Algorithm 7.2 PRIVATESVM-FINITE

Inputs: database $\mathbb{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_{i=1}^{N}$ with $\mathbf{x}^{(i)} \in \Re^D$, $y^{(i)} \in \{-1, 1\}$; finite feature map $\phi : \Re^D \to \Re^F$ and induced kernel k; convex loss function ℓ; and parameters $\lambda, C > 0$.

1 $\boldsymbol{\alpha}^\star \leftarrow$ Run Algorithm 7.1 on $\mathbb{D}$ with parameter C, kernel k, and loss ℓ;
2 $\tilde{\mathbf{w}} \leftarrow \sum_{i=1}^{N} \alpha_i^\star y^{(i)} \phi\left(\mathbf{x}^{(i)}\right)$;
3 $\boldsymbol{\mu} \leftarrow$ Draw i.i.d. sample of F scalars from Laplace $(\mathbf{0}, \lambda)$; and
4 Return $\hat{\mathbf{w}} = \tilde{\mathbf{w}} + \boldsymbol{\mu}$.

Then for every pair of neighboring databases $\mathbb{D}, \mathbb{D}'$ of N entries, we have $\left\|\mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')}\right\|_2 \leq 4LC\kappa/N$, and $\left\|\mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')}\right\|_1 \leq 4LC\kappa\sqrt{F}/N$.

Proof The argument closely follows the proof of the SVM's uniform stability (Schölkopf & Smola 2001, Theorem 12.4). For convenience we define for any training set $\mathbb{S}$

$$R_{\text{reg}}(\mathbf{w}, \mathbb{S}) = \frac{C}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right) + \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$R_{\text{emp}}(\mathbf{w}, \mathbb{S}) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right).$$

Then the first-order necessary KKT conditions imply

$$\mathbf{0} \in \partial_{\mathbf{w}} R_{\text{reg}}(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}) = C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}) + \mathbf{w}^{(\mathbb{D})}, \tag{7.2}$$

$$\mathbf{0} \in \partial_{\mathbf{w}} R_{\text{reg}}(\mathbf{w}^{(\mathbb{D}')}, \mathbb{D}') = C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D}')}, \mathbb{D}') + \mathbf{w}^{(\mathbb{D}')}. \tag{7.3}$$

where $\partial_{\mathbf{w}}$ is the subdifferential operator wrt $\mathbf{w}$. Define the auxiliary risk function

$$\tilde{R}(\mathbf{w}) = C\left\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D}')}, \mathbb{D}'),\ \mathbf{w} - \mathbf{w}^{(\mathbb{D}')} \right\rangle + \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^{(\mathbb{D}')}\right\|_2^2.$$

Note that $\tilde{R}(\cdot)$ maps to sets of reals. It is easy to see that $\tilde{R}(\mathbf{w})$ is strictly convex in $\mathbf{w}$. Substituting $\mathbf{w}^{(\mathbb{D}')}$ into $\tilde{R}(\mathbf{w})$ yields

$$\begin{aligned}\tilde{R}(\mathbf{w}^{(\mathbb{D}')}) &= C\left\langle \partial_{\mathbf{w}} R_{\text{emp}}\left(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}\right) - \partial_{\mathbf{w}} R_{\text{emp}}\left(\mathbf{w}^{(\mathbb{D}')}, \mathbb{D}'\right),\ \mathbf{0} \right\rangle + \frac{1}{2}\|0\|_2^2 \\ &= \{0\},\end{aligned}$$

and by Equation (7.3)

$$\begin{aligned}C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}) + \mathbf{w} &\in C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}) - C\partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D}')}, \mathbb{D}') + \mathbf{w} - \mathbf{w}^{(\mathbb{D}')} \\ &= \partial_{\mathbf{w}} \tilde{R}(\mathbf{w}),\end{aligned}$$

which combined with Equation (7.2) implies $\mathbf{0} \in \partial_{\mathbf{w}} \tilde{R}(\mathbf{w}^{(\mathbb{D})})$, so that $\tilde{R}(\mathbf{w})$ is minimized at $\mathbf{w}^{(\mathbb{D})}$. Thus there exists some nonpositive $r \in \tilde{R}(\mathbf{w}^{(\mathbb{D})})$. Next simplify the first term of $\tilde{R}(\mathbf{w}^{(\mathbb{D})})$, scaled by N/C for notational convenience. In what follows we denote by

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

$\ell'(y, \hat{y})$ the subdifferential $\partial_{\hat{y}} \ell(y, \hat{y})$.

$$
\begin{aligned}
& N \left\langle \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D})}, \mathbb{D}) - \partial_{\mathbf{w}} R_{\text{emp}}(\mathbf{w}^{(\mathbb{D}')}, \mathbb{D}'),\ \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\rangle \\
&= \sum_{i=1}^{N} \left\langle \partial_{\mathbf{w}} \ell\left(y^{(i)}, f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(i)}\right)\right) - \partial_{\mathbf{w}} \ell\left(\hat{y}^{(i)}, f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(i)}\right)\right),\ \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\rangle \\
&= \sum_{i=1}^{N-1} \left(\ell'\left(y^{(i)}, f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(i)}\right)\right) - \ell'\left(y^{(i)}, f_{\mathbf{w}^{(\mathbb{D}')}}\left(\mathbf{x}^{(i)}\right)\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(i)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\mathbf{x}^{(i)}\right)\right) \\
&\quad + \ell'\left(y^{(N)}, f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(N)}\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(N)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\mathbf{x}^{(N)}\right)\right) \\
&\quad - \ell'\left(\hat{y}^{(N)}, f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(N)}\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\hat{\mathbf{x}}^{(N)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(N)}\right)\right) \\
&\geq \ell'\left(y^{(N)}, f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(N)}\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(N)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\mathbf{x}^{(N)}\right)\right) \\
&\quad - \ell'\left(\hat{y}^{(N)}, f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(N)}\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\hat{\mathbf{x}}^{(N)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(N)}\right)\right).
\end{aligned}
$$

Here the second equality follows from $\partial_{\mathbf{w}} \ell(y, f_{\mathbf{w}}(\mathbf{x})) = \ell'(y, f_{\mathbf{w}}(\mathbf{x}))\, \phi(\mathbf{x})$, and $\hat{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)}$ and $\hat{y}^{(i)} = y^{(i)}$ for each $i \in [N-1]$. The inequality follows from the convexity of ℓ in its second argument.[5] Combined with the existence of nonpositive $r \in \tilde{R}(\mathbf{w}^{(\mathbb{D})})$ this yields that there exists

$$
\begin{aligned}
g \in\ & \ell'\left(\hat{y}^{(N)}, f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(N)}\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\hat{\mathbf{x}}^{(N)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\hat{\mathbf{x}}^{(N)}\right)\right) \\
& - \ell'\left(y^{(N)}, f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(N)}\right)\right)\left(f_{\mathbf{w}^{(\mathbb{D})}}\left(\mathbf{x}^{(N)}\right) - f_{\mathbf{w}^{(\mathbb{D}')}}\left(\mathbf{x}^{(N)}\right)\right)
\end{aligned}
$$

such that

$$
\begin{aligned}
0 &\geq \frac{N}{C} r \\
&\geq g + \frac{N}{2C} \left\| \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\|_2^2.
\end{aligned}
$$

And since $|g| \leq 2L \left\| f_{\mathbf{w}^{(\mathbb{D})}} - f_{\mathbf{w}^{(\mathbb{D}')}} \right\|_\infty$ by the Lipschitz continuity of ℓ, this in turn implies

$$
\frac{N}{2C} \left\| \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\|_2^2 \leq 2L \left\| f_{\mathbf{w}^{(\mathbb{D})}} - f_{\mathbf{w}^{(\mathbb{D}')}} \right\|_\infty. \tag{7.4}
$$

Now by the reproducing property and Cauchy-Schwarz inequality we can upper bound the classifier difference's infinity norm by the Euclidean norm on the weight vectors: For each $\mathbf{x}$

$$
\begin{aligned}
\left| f_{\mathbf{w}^{(\mathbb{D})}}(\mathbf{x}) - f_{\mathbf{w}^{(\mathbb{D}')}}(\mathbf{x}) \right| &= \left| \left\langle \phi(\mathbf{x}),\ \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\rangle \right| \\
&\leq \|\phi(\mathbf{x})\|_2 \left\| \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\|_2 \\
&= \sqrt{k(\mathbf{x}, \mathbf{x})} \left\| \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\|_2 \\
&\leq \kappa \left\| \mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')} \right\|_2.
\end{aligned}
$$

[5] Namely for convex f and any $a, b \in \Re$, $(g_a - g_b)(a - b) \geq 0$ for all $g_a \in \partial f(a)$ and all $g_b \in \partial f(b)$.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Combining this with Inequality (7.4) yields $\left\|\mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')}\right\|_2 \leq 4LC\kappa/N$ as claimed. The ℓ_1-based sensitivity then follows from $\|\mathbf{w}\|_1 \leq \sqrt{F}\,\|\mathbf{w}\|_2$ for all $\mathbf{w} \in \Re^F$. □

For a SVM with a Gaussian kernel, we have $L = 1$ and $\kappa = 1$. Then the bounds can be simplified as $\left\|\mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')}\right\|_2 \leq 4C/N$ and $\left\|\mathbf{w}^{(\mathbb{D})} - \mathbf{w}^{(\mathbb{D}')}\right\|_1 \leq 4C\sqrt{F}/N$. With the weight vector's sensitivity in hand, differential privacy follows immediately (cf. Lemma 7.3)

THEOREM 7.10 (Privacy of PRIVATESVM-FINITE) *For any $\beta > 0$, database $\mathbb{D}$ of size N, $C > 0$, loss function $\ell(y, \hat{y})$ that is convex and L-Lipschitz in $\hat{y}$, and finite F-dimensional feature map with kernel $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2$ for all $\mathbf{x} \in \Re^D$,* PRIVATESVM-FINITE *run on $\mathbb{D}$ with loss ℓ, kernel k, noise parameter $\lambda \geq 4LC\kappa\sqrt{F}/(\beta N)$, and regularization parameter C guarantees β-differential privacy.*

This result states that higher levels of privacy require more noise, while more training examples reduce the level of required noise. We next establish the (ϵ, δ)-usefulness of PRIVATESVM-FINITE, using the exponential tails of the noise vector $\boldsymbol{\mu}$. By contrast to privacy, utility demands that the noise not be too large.

THEOREM 7.11 (Utility of PRIVATESVM-FINITE) *Consider any $C > 0$, $N > 1$, database $\mathbb{D}$ of N entries, arbitrary convex loss ℓ, and finite F-dimensional feature mapping ϕ with kernel k and $\left|\phi(\mathbf{x})_i\right| \leq \Phi$ for all $\mathbf{x} \in \mathcal{M}$ and $i \in [F]$ for some $\Phi > 0$ and $\mathcal{M} \subseteq \Re^D$. For any $\epsilon > 0$ and $\delta \in (0, 1)$,* PRIVATESVM-FINITE *run on $\mathbb{D}$ with loss ℓ, kernel k, noise parameter $0 < \lambda \leq \frac{\epsilon}{2\Phi\left(F + \log_e \frac{1}{\delta}\right)}$, and regularization parameter C is (ϵ, δ)-useful with respect to the* SVM *under the $\|\cdot\|_{\infty;\mathcal{M}}$-norm.*

In other words, run with arbitrary noise parameter $\lambda > 0$, PRIVATESVM-FINITE is (ϵ, δ)-useful for $\epsilon = \Omega\left(\lambda\Phi\left(F + \log_e \frac{1}{\delta}\right)\right)$.

Proof Consider the SVM and PRIVATESVM-FINITE classifications on an arbitrary point $\mathbf{x} \in \mathcal{M}$:

$$\begin{aligned}\left|f_{\hat{M}(\mathbb{D})}(\mathbf{x}) - f_{M(\mathbb{D})}(\mathbf{x})\right| &= \left|\langle\hat{\mathbf{w}},\, \phi(\mathbf{x})\rangle - \langle\tilde{\mathbf{w}},\, \phi(\mathbf{x})\rangle\right| \\ &= \left|\langle\boldsymbol{\mu},\, \phi(\mathbf{x})\rangle\right| \\ &\leq \|\boldsymbol{\mu}\|_1\, \|\phi(\mathbf{x})\|_\infty \\ &\leq \Phi\,\|\boldsymbol{\mu}\|_1 .\end{aligned}$$

The absolute value of a zero-mean Laplace random variable with scale λ is exponentially distributed with scale λ^{-1}. Moreover, the sum of q i.i.d. exponential random variables has an Erlang q-distribution with the same scale parameter. Thus we have, for Erlang F-distributed random variable X and any $t > 0$,

$$\begin{aligned}&\forall\mathbf{x} \in \mathcal{M},\ \left|f_{\hat{M}(\mathbb{D})}(\mathbf{x}) - f_{M(\mathbb{D})}(\mathbf{x})\right| \leq \Phi X \\ \Rightarrow\ \forall\epsilon > 0,\ \Pr\left(\left\|f_{\hat{M}(\mathbb{D})} - f_{M(\mathbb{D})}\right\|_{\infty;\mathcal{M}} > \epsilon\right) &\leq \Pr(X > \epsilon/\Phi) \\ &\leq \frac{\mathrm{E}\left[e^{tX}\right]}{e^{\epsilon t/\Phi}}. \qquad (7.5)\end{aligned}$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Algorithm 7.3 PRIVATEERM-OBJECTIVE

Inputs: database $\mathbb{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_{i=1}^{N}$ with $\mathbf{x}^{(i)} \in \Re^D$, $y^{(i)} \in \{-1, 1\}$; regularizer $\rho(\cdot)$; loss function $\ell(\cdot)$; $c > 0$ a bound on the 2nd derivative of ℓ; regularization parameter $\lambda > 0$; and privacy parameter $\beta > 0$.

1 Let $\beta' = \beta - \log\left(1 + \frac{2c}{N\lambda} + \frac{c^2}{N^2\lambda^2}\right)$;
2 If $\beta' > 0$, then let $\Delta = 0$; otherwise let $\Delta = \frac{c}{N(e^{\beta/4}-1)} - \lambda$ and let $\beta' = \beta/2$;
3 Sample $\mathbf{b} \sim \exp(-\beta'\|\mathbf{b}\|_2/2)$; and
4 Return $\mathbf{f}_{priv} = \arg\min_{\mathbf{f}} J_{priv}(\mathbf{f}, \mathbb{D}) + \frac{1}{2}\Delta\|\mathbf{f}\|_2^2$.

Here we have employed the Chernoff tail bound technique using Markov's inequality. The numerator of (7.5), the moment-generating function of the Erlang F-distribution with parameter λ, is $(1 - \lambda t)^{-F}$ for $t < \lambda^{-1}$. With the choice of $t = (2\lambda)^{-1}$, this gives

$$\begin{aligned}\Pr\left(\left\|f_{\hat{M}(\mathbb{D})} - f_{M(\mathbb{D})}\right\|_{\infty;\mathcal{M}} > \epsilon\right) &\leq (1 - \lambda t)^{-F} e^{-\epsilon t/\Phi} \\ &= 2^F e^{-\epsilon/(2\lambda\Phi)} \\ &= \exp\left(F \log_e 2 - \frac{\epsilon}{2\lambda\Phi}\right) \\ &< \exp\left(F - \frac{\epsilon}{2\lambda\Phi}\right).\end{aligned}$$

And provided that $\epsilon \geq \left(2\lambda\Phi\left(F + \log_e \frac{1}{\delta}\right)\right)$ this probability is bounded by δ. □

7.5 Differential Privacy by Objective Perturbation

We briefly review another early independent approach to differentially private SVM learning due to Chaudhuri, Monteleoni, & Sarwate (2011). Their mechanism for linear SVM guarantees differential privacy by adding a random term to the objective, as performed previously by the same group for regularized logistic regression (Chaudhuri & Monteleoni 2009) and as is suitable for a general class of regularized empirical risk minimizers (Chaudhuri et al. 2011). While we limit our discussion to the linear SVM for clarity of exposition, all results go through essentially unchanged for SVM under finite-dimensional feature mappings. It is also notable that further general connections and treatments of (regularized) empirical risk minimization and differential privacy have been made; for example, with Bassily et al. (2014) providing matching rates under bounded Lipschitz influence of each training datum on loss and bounded parameter domains; and Duchi et al. (2013) demonstrating minimax rates for efficient convex risk minimization under a definition of local privacy (when data is unavailable to the statistician) with lower/upper bounds matching up to constant factors.

We repeat here the objective of regularized empirical risk minimization, found in Equation (2.1):

$$J(f, \mathbb{D}) = \frac{1}{N} \sum_{(x,y)\in\mathbb{D}} \ell\left(y, f(x)\right) + \lambda \cdot \rho\left(f\right),$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

where $\rho(\cdot)$ is a regularization functional that for the SVM is one-half the ℓ_2-norm, and where $\lambda > 0$ is the regularization parameter equivalent to $1/C$ for the SVM as introduced earlier. The mechanism PrivateERM-Objective, as described by Algorithm 7.3, aims to minimize

$$J_{priv}(\mathbf{f}, \mathbb{D}) = J(\mathbf{f}, \mathbb{D}) + \frac{1}{N}\langle \mathbf{b}, \mathbf{f} \rangle,$$

where $\mathbf{b}$ is zero-mean random noise. Intuitively, the perturbed objective minimizes regularized empirical risk while favoring models that align with a random direction. The strength of alignment preferred—the degree of objective perturbation—depends on the level of privacy required. To provide a flavor of the results known for PrivateERM-Objective, we state privacy and utility guarantees due to Chaudhuri et al. (2011), directing the interested reader to the original paper for detailed discussion and proofs.

THEOREM 7.12 (Privacy of PrivateERM-Objective; Chaudhuri et al. 2011, Theorem 6) *If regularization functional $\rho(\cdot)$ is 1-strongly convex and doubly differentiable, and loss function $\ell(\cdot)$ is convex and doubly differentiable with first and second derivatives bounded by 1 and $c > 0$, respectively, then* PrivateERM-Objective *is β-differentially private.*

It is noteworthy that preserving privacy via the randomized objective can only apply to convex differentiable loss functions, ruling out the most common case for the SVM: the nondifferentiable hinge loss. The output perturbation approach of the previous section preserves privacy for any convex loss—a very weak condition since convexity is required for the formulation of SVM learning to be convex. Chaudhuri et al. (2011) explore two instantiations of PrivateERM-Objective for the SVM with this limitation in mind. The more complex approach is to use the Huber loss, which is not globally doubly differentiable, but nonetheless can be shown to achieve differential privacy. The simpler alternative is to use the following loss function that satisfies the conditions of Theorem 7.12 with $c = \frac{3}{4h}$.

$$\ell_s(z) = \begin{cases} 0, & \text{if } z > 1 + h \\ -\frac{(1-z)^4}{16h^3} + \frac{3(1-z)^2}{8h} + \frac{1-z}{2} + \frac{3h}{16}, & \text{if } |1 - z| \leq h. \\ 1 - z, & \text{if } z < 1 - h \end{cases}$$

As the bandwidth $h \to 0$, this loss approaches the hinge loss. And using this loss with resulting $c = \frac{3}{4h}$, and regularization functional $\rho(\cdot) = \frac{1}{2}\|\mathbf{f}\|_2^2$, PrivateERM-Objective yields a β-differentially private approximation of the SVM (Chaudhuri et al. 2011, Corollary 12).

While the definition of utility of Section 7.2.2 measures the pointwise similarity of the private SVM classifier to the nonprivate SVM classifier, Chaudhuri et al. (2011) measure utility in terms of bounds on excess risk.

THEOREM 7.13 (Excess Risk of PrivateERM-Objective; (Chaudhuri et al. 2011), Theorem 18) *Consider regularization functional $\rho(\mathbf{f}) = \frac{1}{2}\|\mathbf{f}\|_2^2$, D-dimensional data $\mathbb{D}$ drawn i.i.d. according to distribution $P_{\mathcal{Z}}$, and $\mathbf{f}_0$ a reference classifier with some risk $R(P_{\mathcal{Z}}, \mathbf{f}) = R^\star$. Under the same conditions as in Theorem 7.12, there exists a constant*

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

$A > 0$ such that for $\delta > 0$, if the training set size satisfies

$$N > A \cdot \max\left\{\frac{\|\mathbf{f}_0\|_2^2 \log(1/\delta)}{\epsilon^2}, \frac{c\|\mathbf{f}_0\|_2^2}{\epsilon\beta}, \frac{D\log\left(\frac{D}{\delta}\right)\|\mathbf{f}_0\|_2^2}{\epsilon\beta}\right\}.$$

then the excess risk of PrivateERM-Objective *is bounded with high probability*

$$\Pr\left(R\left(P_{\mathcal{Z}}, \mathbf{f}_{priv}\right) \leq R^\star + \epsilon\right) \geq 1 - 2\delta.$$

For the purpose of comparison, nonprivate SVM requires data size of at least a constant times $\|\mathbf{f}_0\|_2^2 \log(1/\delta)/\epsilon^2$ to achieve the same guarantee on excess risk (Shalev-Shwartz & Srebro 2008). This corresponds to the first term of the max in the sample complexity of PrivateERM-Objective.

It is noteworthy that for SVM learning with the hinge loss, guarantees on pointwise similarity utility are strictly stronger than risk bounds.

Remark 7.14 Since the hinge loss is Lipschitz in the classifier output by the SVM, any mechanism $\hat{M}$ having utility with respect to the SVM also has expected hinge loss that is within ϵ of the SVM's hinge loss with high probability; i.e., (ϵ, δ)-usefulness with respect to the sup-norm is stronger than guaranteed closeness of risk.

The stronger definition of utility offers a natural advantage: An arbitrary differentially private mechanism that enjoys low risk is not necessarily an approximation of a given learning algorithm of interest; it is natural to expect that a private SVM approximates the classifications of a nonprivate SVM. Guarantees with respect to this utility imply such approximation and (for the SVM) low risk.

While analytical results for PrivateSVM-Finite and PrivateERM-Objective are not directly comparable, the excess risk bounds for PrivateERM-Objective enjoy better growth rates than those proved by Chaudhuri et al. (2011) for output perturbation, and early experiments on benchmark datasets suggest that objective perturbation can outperform output perturbation.

Finally, Chaudhuri et al. (2011) also develop a method for tuning the regularization parameter while preserving privacy, using a comparison procedure due to McSherry & Talwar (2007).

For the remainder of this chapter, we focus on the mechanism of output perturbation.

7.6 Infinite-Dimensional Feature Spaces

We now consider the problem of privately learning in an RKHS $\mathcal{H}$ induced by an infinite-dimensional feature mapping ϕ. We begin the section by deriving the mechanism, then establish the range of noise parameters required to guarantee privacy (Corollary 7.15), and derive the conditions under which the mechanism yields close approximations to the nonprivate SVM (Theorem 7.16).

It is natural to look to the dual SVM as a starting point: an optimizing $f^\star \in \mathcal{H}$ must lie in the span of the data by the Representer Theorem (Kimeldorf & Wahba 1971). While the coordinates with respect to this data basis—the $\alpha_i^\star$ dual variables—could be perturbed to guarantee differential privacy, the basis is also needed to parametrize $f^\star$. The

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Algorithm 7.4 PrivateSVM

Inputs: database $\mathbb{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_{i=1}^{N}$ with $\mathbf{x}^{(i)} \in \Re^D$, $y^{(i)} \in \{-1, 1\}$; translation-invariant kernel $k(\mathbf{x}, \mathbf{z}) = g(\mathbf{x} - \mathbf{z})$ with Fourier transform $p(\boldsymbol{\omega}) = 2^{-1} \int e^{-j\langle \boldsymbol{\omega}, \mathbf{x} \rangle} g(\mathbf{x})\, d\mathbf{x}$; convex loss function ℓ; parameters $\lambda, C > 0$, and $\hat{D} \in \mathfrak{N}$.

1 $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{\hat{D}} \leftarrow$ Draw i.i.d. sample of $\hat{D}$ vectors in $\Re^D$ from p;
2 $\hat{\boldsymbol{\alpha}} \leftarrow$ Run Algorithm 7.1 on $\mathbb{D}$ with parameter C, kernel $\hat{k}$ induced by map (7.6), and loss ℓ;
3 $\tilde{\mathbf{w}} \leftarrow \sum_{i=1}^{N} y^{(i)} \hat{\alpha}_i \hat{\phi}\left(\mathbf{x}^{(i)}\right)$ where $\hat{\phi}$ is defined in Equation (7.6);
4 $\boldsymbol{\mu} \leftarrow$ Draw i.i.d. sample of $2\hat{D}$ scalars from Laplace $(\mathbf{0}, \lambda)$; and
5 Return $\hat{\mathbf{w}} = \tilde{\mathbf{w}} + \boldsymbol{\mu}$ and $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{\hat{D}}$.

basis is the original data itself, so such an approach appears to be a dead end. Instead we approach the problem by approximating $\mathcal{H}$ with a random RKHS $\hat{\mathcal{H}}$ induced by a random finite-dimensional map $\hat{\phi}$, which admits a response based on a primal parametrization. This idea was applied independently by both Chaudhuri et al. (2011) and Rubinstein et al. (2012) to the output- and objective-perturbation mechanisms. Algorithm 7.4 summarizes this mechanism from Rubinstein (2010).

As noted by Rahimi & Recht (2008), the Fourier transform p of the kernel function g, a continuous positive-definite translation-invariant function, is a non-negative measure (Rudin 1994). If the kernel g is properly scaled, Bochner's theorem guarantees that p is a proper probability distribution. Rahimi & Recht (2008) exploit this fact to construct a random RKHS $\hat{\mathcal{H}}$ by drawing $\hat{D}$ vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{\hat{D}}$ from p, and defining the random $2\hat{D}$-dimensional feature map

$$\hat{\phi}(\cdot) = \hat{D}^{-1/2} \left[\cos\left(\langle \boldsymbol{\rho}_1, \cdot \rangle\right), \sin\left(\langle \boldsymbol{\rho}_1, \cdot \rangle\right), \ldots, \cos\left(\langle \boldsymbol{\rho}_{\hat{D}}, \cdot \rangle\right), \sin\left(\langle \boldsymbol{\rho}_{\hat{D}}, \cdot \rangle\right)\right]^T. \tag{7.6}$$

Table 7.2 presents three translation-invariant kernels and their transformations. Inner products in the random feature space $\hat{k}(\cdot, \cdot)$ approximate $k(\cdot, \cdot)$ uniformly and arbitrary precision depending on parameter $\hat{D}$, as restated in Lemma 7.21. Rahimi & Recht (2008) apply this approximation to large-scale learning, finding good approximations for $\hat{D} \ll N$. We perform regularized ERM in $\hat{\mathcal{H}}$, not to avoid complexity in N, but to provide a direct finite representation $\tilde{\mathbf{w}}$ of the primal solution in the case of infinite-dimensional feature spaces. Subsequently, Laplace noise is added to the primal solution $\tilde{\mathbf{w}}$ to guarantee differential privacy as before.

Unlike PrivateSVM-Finite, PrivateSVM must release a parametrization of feature map $\hat{\phi}$—the sample $\left\{\boldsymbol{\rho}_i\right\}_{i=1}^{\hat{D}}$—to classify as $\hat{f}^\star = \langle \hat{\mathbf{w}}, \hat{\phi}(\cdot) \rangle$. Of PrivateSVM's response, only $\hat{\mathbf{w}}$ depends on $\mathbb{D}$; the $\boldsymbol{\rho}_i$ are data-independent draws from the kernel's transform p, which we assume to be known by the adversary (to wit the adversary knows the mechanism, including k). Thus to establish differential privacy we need only consider the weight vector, as we did for PrivateSVM-Finite.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Table 7.2 Translation-invariant kernels of Table 7.1, their g functions, and the corresponding Fourier transforms p

Kernel	$g(\boldsymbol{\Delta})$	$p(\boldsymbol{\omega})$
RBF	$\exp\left(-\frac{\|\boldsymbol{\Delta}\|_2^2}{2\sigma^2}\right)$	$\frac{1}{(2\pi)^{D/2}}\exp\left(\frac{-\|\boldsymbol{\omega}\|_2^2}{2}\right)$
Laplacian	$\exp\left(-\|\boldsymbol{\Delta}\|_1\right)$	$\prod_{i=1}^{D}\frac{1}{\pi\left(1+\omega_i^2\right)}$
Cauchy	$\prod_{i=1}^{D}\frac{2}{1+\Delta_i^2}$	$\exp\left(-\|\boldsymbol{\omega}\|_1\right)$

COROLLARY 7.15 (Privacy of PRIVATESVM) *For any $\beta > 0$, database $\mathbb{D}$ of size N, $C > 0$, $\hat{D} \in \mathfrak{N}$, loss function $\ell(y, \hat{y})$ that is convex and L-Lipschitz in $\hat{y}$, and translation-invariant kernel k,* PRIVATESVM *run on $\mathbb{D}$ with loss ℓ, kernel k, noise parameter $\lambda \geq 2^{2.5} LC\sqrt{\hat{D}}/(\beta N)$, approximation parameter $\hat{D}$, and regularization parameter C guarantees β-differential privacy.*

Proof The result follows from Theorem 7.10 since $\tilde{\mathbf{w}}$ is the primal solution of SVM with kernel $\hat{k}$, the response vector $\hat{\mathbf{w}} = \tilde{\mathbf{w}} + \boldsymbol{\mu}$, and $\hat{k}(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x} \in \Re^D$. The extra factor of $\sqrt{2}$ comes from the fact that $\hat{\phi}(\cdot)$ is a $2\hat{d}$-dimensional feature map. □

This result is surprising, in that PRIVATESVM is able to guarantee privacy for regularized ERM over a function class of infinite dimension, where the obvious way to return the learned classifier (responding with the dual variables and feature mapping) reveals all the entries corresponding to the support vectors completely.

The remainder of this section considers the following result, which states that PRIVATESVM is useful with respect to the SVM.

THEOREM 7.16 (Utility of PRIVATESVM) *Consider any database $\mathbb{D}$, compact set $\mathcal{M} \subseteq \Re^D$ containing $\mathbb{D}$, convex loss ℓ, translation-invariant kernel k, and scalars $C, \epsilon > 0$ and $\delta \in (0, 1)$. Suppose the SVM with loss ℓ, kernel k, and parameter C has dual variables with ℓ_1-norm bounded by Λ. Then Algorithm 7.4 run on $\mathbb{D}$ with loss ℓ, kernel k, parameters $\hat{D} \geq \frac{4(D+2)}{\theta(\epsilon)}\log_e\left(\frac{2^9\left(\sigma_p \mathrm{diam}(\mathcal{M})\right)^2}{\delta\theta(\epsilon)}\right)$ where $\theta(\epsilon) = \min\left\{1, \frac{\epsilon^4}{2^4\left(\Lambda+2\sqrt{(CL+\Lambda/2)\Lambda}\right)^4}\right\}$, $\lambda \leq \min\left\{\frac{\epsilon}{2^4\log_e 2\sqrt{\hat{D}}}, \frac{\epsilon\sqrt{\hat{D}}}{8\log_e\frac{2}{\delta}}\right\}$, and C is (ϵ, δ)-useful with respect to Algorithm 7.1 run on $\mathbb{D}$ with loss ℓ, kernel k and parameter C, with respect to the $\|\cdot\|_{\infty;\mathcal{M}}$-norm.*

Remark 7.17 Theorem 7.16 introduces the assumption that the SVM has a dual solution vector with bounded ℓ_1-norm. The motivation for this condition is the most common case for SVM classification: learning with the hinge loss. Under this loss the dual program (7.1) has box constraints that ensure that this condition is satisfied.

The result of Theorem 7.16 bounds the pointwise distance between classifiers $f^\star$ output by SVM and $\hat{f}^\star$ output by PRIVATESVM with high probability. Let $\tilde{f}$ be the function parametrized by intermediate-weight vector $\tilde{\mathbf{w}}$. Then we establish the main result

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

by proving that both $f^\star$ and $\hat{f}^\star$ are close to $\tilde{f}$ with high probability and applying the triangle inequality. We begin by relating $\tilde{f}$ and $f^\star$. As $f^\star$ is the result of adding Laplace noise to $\tilde{\mathbf{w}}$, the task of relating these two classifiers is almost the same as proving the utility of PrivateSVM-Finite (see Theorem 7.11).

COROLLARY 7.18 *Consider a run of Algorithms 7.1 and 7.4 with $\hat{D} \in \mathfrak{N}$, $C > 0$, convex loss, and translation-invariant kernel. Denote by $\hat{f}^\star$ and $\tilde{f}$ the classifiers parametrized by weight vectors $\hat{\mathbf{w}}$ and $\tilde{\mathbf{w}}$, respectively, where these vectors are related by $\hat{\mathbf{w}} = \tilde{\mathbf{w}} + \boldsymbol{\mu}$ with $\boldsymbol{\mu} \overset{iid}{\sim} \text{Laplace}(\mathbf{0}, \lambda)$ in Algorithm 7.4. For any $\epsilon > 0$ and $\delta \in (0, 1)$, if $0 < \lambda \leq \min\left\{\frac{\epsilon}{2^4 \log_e 2\sqrt{\hat{D}}}, \frac{\epsilon\sqrt{\hat{D}}}{8 \log_e \frac{2}{\delta}}\right\}$ then $\Pr\left(\left\|\hat{f}^\star - \tilde{f}\right\|_\infty \leq \frac{\epsilon}{2}\right) \geq 1 - \frac{\delta}{2}$.*

Proof As in the proof of Theorem 7.11 we can use the Chernoff trick to show that, for an Erlang $2\hat{D}$-distributed random variable X, the choice of $t = (2\lambda)^{-1}$, and for any $\epsilon > 0$

$$\begin{aligned}
\Pr\left(\left\|\hat{f}^\star - \tilde{f}\right\|_\infty > \epsilon/2\right) &\leq \frac{\mathrm{E}\left[e^{tX}\right]}{e^{\epsilon t\sqrt{\hat{D}}/2}} \\
&\leq (1 - \lambda t)^{-2\hat{D}}\, e^{-\epsilon t\sqrt{\hat{D}}/2} \\
&= 2^{2\hat{D}} e^{-\epsilon\sqrt{\hat{D}}/(4\lambda)} \\
&= \exp\left(\hat{D}\log_e 4 - \epsilon\sqrt{\hat{D}}/(4\lambda)\right).
\end{aligned}$$

Provided that $\lambda \leq \epsilon / \left(2^4 \log_e 2\sqrt{\hat{D}}\right)$ this is bounded by $\exp\left(-\epsilon\sqrt{\hat{D}}/(8\lambda)\right)$. Moreover, if $\lambda \leq \epsilon\sqrt{\hat{D}} / \left(8 \log_e \frac{2}{\delta}\right)$, then the claim follows. □

To relate $f^\star$ and $\tilde{f}$, we exploit smoothness of regularized ERM with respect to small changes in the RKHS itself. We begin with a technical lemma that we will use to exploit the convexity of the regularized empirical risk functional; it shows a kind of converse to Remark 7.14 relating that functions with risks that are close in value will also be close in proximity.

LEMMA 7.19 *Let R be a functional on Hilbert space $\mathcal{H}$ satisfying $R[f] \geq R[f^\star] + \frac{a}{2}\left\|f - f^\star\right\|_{\mathcal{H}}^2$ for some $a > 0$, $f^\star \in \mathcal{H}$, and all $f \in \mathcal{H}$. Then $R[f] \leq R[f^\star] + \epsilon$ implies $\left\|f - f^\star\right\|_{\hat{\mathcal{H}}} \leq \sqrt{\frac{2\epsilon}{a}}$, for all $\epsilon > 0$, $f \in \mathcal{H}$.*

Proof By assumption and the antecedent

$$\begin{aligned}
\left\|f - f^\star\right\|_{\hat{\mathcal{H}}}^2 &\leq \frac{2}{a}\left(R[f] - R[f^\star]\right) \\
&\leq \frac{2}{a}\left(R[f^\star] + \epsilon - R[f*]\right) \\
&= \frac{2\epsilon}{a}.
\end{aligned}$$

Taking square roots of both sides yields the result. □

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Provided that the kernels $k, \hat{k}$ are uniformly close, we now show that $f^\star$ and $\tilde{f}$ are pointwise close, using the insensitivity of regularized ERM to feature mapping perturbation.

LEMMA 7.20 *Let $\mathcal{H}$ be an RKHS with translation-invariant kernel k, and let $\hat{\mathcal{H}}$ be the random RKHS corresponding to feature map (7.6) induced by k. Let C be a positive scalar and loss $\ell(y, \hat{y})$ be convex and L-Lipschitz continuous in $\hat{y}$. Consider the regularized empirical risk minimizers in each RKHS, where $R_{\text{emp}}[f] = n^{-1} \sum_{i=1}^{N} \ell\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right)$,*

$$f^\star \in \operatorname*{argmin}_{f \in \mathcal{H}} \left[C R_{\text{emp}}[f] + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right]$$
$$g^\star \in \operatorname*{argmin}_{g \in \hat{\mathcal{H}}} \left[C R_{\text{emp}}[g] + \frac{1}{2} \|g\|_{\hat{\mathcal{H}}}^2 \right].$$

Let $\mathcal{M} \subseteq \Re^D$ be any set containing $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$. For any $\epsilon > 0$, if the dual variables from both optimizations have ℓ_1-norms bounded by some $\Lambda > 0$ and $\left\| k - \hat{k} \right\|_{\infty;\mathcal{M}} \leq \min\left\{1, \frac{\epsilon^2}{2^2\left(\Lambda + 2\sqrt{(CL+\Lambda/2)\Lambda}\right)^2}\right\}$ then $\left\| f^\star - g^\star \right\|_{\infty;\mathcal{M}} \leq \epsilon/2$.

Proof Define regularized empirical risk functional $R_{\text{reg}}[f] = C R_{\text{emp}}[f] + \|f\|^2/2$, for the appropriate RKHS norm. Let minimizer $f^\star \in \mathcal{H}$ be given by parameter vector $\boldsymbol{\alpha}^\star$, and let minimizer $g^\star \in \hat{\mathcal{H}}$ be given by parameter vector $\boldsymbol{\beta}^\star$. Let $g_{\boldsymbol{\alpha}^\star} = \sum_{i=1}^{N} \alpha_i^\star y^{(i)} \hat{\phi}\left(\mathbf{x}^{(i)}\right)$ and $f_{\boldsymbol{\beta}^\star}^\star = \sum_{i=1}^{N} \beta_i^\star y^{(i)} \phi\left(\mathbf{x}^{(i)}\right)$ denote the images of $f^\star$ and $g^\star$ under the natural mapping between the spans of the data in RKHS's $\hat{\mathcal{H}}$ and $\mathcal{H}$, respectively. We will first show that these four functions have arbitrarily close regularized empirical risk in their respective RKHS, and then that this implies uniform proximity of the functions themselves. Observe that for any $g \in \hat{\mathcal{H}}$,

$$\begin{aligned} R_{\text{reg}}^{\hat{\mathcal{H}}}[g] &= C R_{\text{emp}}[g] + \frac{1}{2} \|g\|_{\hat{\mathcal{H}}}^2 \\ &\geq C \langle \partial_g R_{\text{emp}}[g^\star],\, g - g^\star \rangle_{\hat{\mathcal{H}}} + C R_{\text{emp}}[g^\star] + \frac{1}{2} \|g\|_{\hat{\mathcal{H}}}^2 \\ &= \langle \partial_g R_{\text{reg}}^{\hat{\mathcal{H}}}[g^\star],\, g - g^\star \rangle_{\hat{\mathcal{H}}} - \langle g^\star,\, g - g^\star \rangle_{\hat{\mathcal{H}}} + C R_{\text{emp}}[g^\star] + \frac{1}{2} \|g\|_{\hat{\mathcal{H}}}^2 . \end{aligned}$$

The inequality follows from the convexity of $R_{\text{emp}}[\cdot]$ and holds for all elements of the subdifferential $\partial_g R_{\text{emp}}[g^\star]$. The subsequent equality holds by $\partial_g R_{\text{reg}}^{\hat{\mathcal{H}}}[g] = C\, \partial_g R_{\text{emp}}[g] + g$. Now since $\mathbf{0} \in \partial_g R_{\text{reg}}^{\hat{\mathcal{H}}}[g^\star]$, it follows that

$$\begin{aligned} R_{\text{reg}}^{\hat{\mathcal{H}}}[g] &\geq C R_{\text{emp}}[g^\star] + \frac{1}{2} \|g\|_{\hat{\mathcal{H}}}^2 - \langle g^\star,\, g - g^\star \rangle_{\hat{\mathcal{H}}} \\ &= R_{\text{reg}}^{\hat{\mathcal{H}}}[g^\star] + \frac{1}{2} \|g\|_{\hat{\mathcal{H}}}^2 - \langle g^\star,\, g \rangle_{\hat{\mathcal{H}}} + \frac{1}{2} \left\|g^\star\right\|_{\hat{\mathcal{H}}}^2 \\ &= R_{\text{reg}}^{\hat{\mathcal{H}}}[g^\star] + \frac{1}{2} \left\|g - g^\star\right\|_{\hat{\mathcal{H}}}^2 . \end{aligned}$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

With this, Lemma 7.19 states that for any $g \in \hat{\mathcal{H}}$ and $\epsilon' > 0$,

$$R^{\hat{\mathcal{H}}}_{\text{reg}}[g] \leq R^{\hat{\mathcal{H}}}_{\text{reg}}[g^\star] + \epsilon' \;\Rightarrow\; \left\|g - g^\star\right\|_{\hat{\mathcal{H}}} \leq \sqrt{2\epsilon'}. \tag{7.7}$$

Next we show that the antecedent is true for $g = g_{\boldsymbol{\alpha}^\star}$. Conditioned on $\left\{\left\|k - \hat{k}\right\|_{\infty;\mathcal{M}} \leq \epsilon'\right\}$, for all $\mathbf{x} \in \mathcal{M}$

$$\begin{aligned}
\left|f^\star_{[}](\mathbf{x}) - g_{\boldsymbol{\alpha}^\star}(\mathbf{x})\right| &= \left|\sum_{i=1}^{N} \alpha_i^\star y^{(i)} \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}\right) - \hat{k}\left(\mathbf{x}^{(i)}, \mathbf{x}\right)\right)\right| \\
&\leq \sum_{i=1}^{N} \left|\alpha_i^\star\right| \left|k\left(\mathbf{x}^{(i)}, \mathbf{x}\right) - \hat{k}\left(\mathbf{x}^{(i)}, \mathbf{x}\right)\right| \\
&\leq \epsilon' \left\|\boldsymbol{\alpha}^\star\right\|_1 \\
&\leq \epsilon' \Lambda,
\end{aligned} \tag{7.8}$$

by the bound on $\|\boldsymbol{\alpha}^\star\|_1$. This and the Lipschitz continuity of the loss lead to

$$\begin{aligned}
\left|R^{\mathcal{H}}_{\text{reg}}[f^\star] - R^{\hat{\mathcal{H}}}_{\text{reg}}[g_{\boldsymbol{\alpha}^\star}]\right| &= \left|C\, R_{\text{emp}}[f^\star] - C\, R_{\text{emp}}[g_{\boldsymbol{\alpha}^\star}] + \frac{1}{2}\left\|f^\star\right\|^2_{\mathcal{H}} - \frac{1}{2}\left\|g_{\boldsymbol{\alpha}^\star}\right\|^2_{\hat{\mathcal{H}}}\right| \\
&\leq \frac{C}{N} \sum_{i=1}^{N} \left|\ell\left(y^{(i)}, f^\star_{[}\left(\mathbf{x}^{(i)}\right)\right) - \ell\left(y^{(i)}, g_{\boldsymbol{\alpha}^\star}(\mathbf{x}^{(i)})\right)\right| \\
&\quad + \frac{1}{2}\left|(\boldsymbol{\alpha}^\star)^\top \left(\mathbf{K} - \hat{\mathbf{K}}\right)\boldsymbol{\alpha}^\star\right| \\
&\leq CL \left\|f^\star - g_{\boldsymbol{\alpha}^\star}\right\|_{\infty;\mathcal{M}} + \frac{1}{2}\left\|\boldsymbol{\alpha}^\star\right\|_1 \left\|\left(\mathbf{K} - \hat{\mathbf{K}}\right)\boldsymbol{\alpha}^\star\right\|_\infty \\
&\leq CL\epsilon'\Lambda + \Lambda^2\epsilon'/2 \\
&= \left(CL + \frac{\Lambda}{2}\right)\Lambda\epsilon',
\end{aligned}$$

where $\mathbf{K}$ and $\hat{\mathbf{K}}$ are the kernel matrices of the kernels k and $\hat{k}$, respectively. Similarly,

$$\left|R^{\hat{\mathcal{H}}}_{\text{reg}}[g^\star] - R^{\mathcal{H}}_{\text{reg}}[f_{\boldsymbol{\beta}^\star}]\right| \leq (CL + \Lambda/2)\Lambda\epsilon'$$

by the same argument. And since $R^{\mathcal{H}}_{\text{reg}}[f_{\boldsymbol{\beta}^\star}] \geq R^{\mathcal{H}}_{\text{reg}}[f^\star]$ and $R^{\hat{\mathcal{H}}}_{\text{reg}}[g_{\boldsymbol{\alpha}^\star}] \geq R^{\hat{\mathcal{H}}}_{\text{reg}}[g^\star]$, we have proved that

$$\begin{aligned}
R^{\hat{\mathcal{H}}}_{\text{reg}}[g_{\boldsymbol{\alpha}^\star}] &\leq R^{\mathcal{H}}_{\text{reg}}[f^\star] + (CL + \Lambda/2)\Lambda\epsilon' \\
&\leq R^{\mathcal{H}}_{\text{reg}}[f_{\boldsymbol{\beta}^\star}] + (CL + \Lambda/2)\Lambda\epsilon' \\
&\leq R^{\hat{\mathcal{H}}}_{\text{reg}}[g^\star] + 2(CL + \Lambda/2)\Lambda\epsilon'.
\end{aligned}$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

And by implication (7.7),

$$\left\|g_{\boldsymbol{\alpha}^\star} - g^\star\right\|_{\hat{\mathcal{H}}} \leq 2\sqrt{\left(CL + \frac{\Lambda}{2}\right)\Lambda\epsilon'}. \tag{7.9}$$

Now $\hat{k}(\mathbf{x}, \mathbf{x}) = 1$ for each $\mathbf{x} \in \Re^D$ implies

$$\begin{aligned}\left|g_{\boldsymbol{\alpha}^\star}(\mathbf{x}) - g^\star(\mathbf{x})\right| &= \left\langle g_{\boldsymbol{\alpha}^\star} - g^\star,\ \hat{k}(\mathbf{x}, \cdot)\right\rangle_{\hat{\mathcal{H}}} \\ &\leq \left\|g_{\boldsymbol{\alpha}^\star} - g^\star\right\|_{\hat{\mathcal{H}}} \sqrt{\hat{k}(\mathbf{x}, \mathbf{x})} \\ &= \left\|g_{\boldsymbol{\alpha}^\star} - g^\star\right\|_{\hat{\mathcal{H}}}.\end{aligned}$$

This combines with Inequality (7.9) to yield $\left\|g_{\boldsymbol{\alpha}^\star} - g^\star\right\|_{\infty;\mathcal{M}} \leq 2\sqrt{\left(CL + \frac{\Lambda}{2}\right)\Lambda\epsilon'}$. Together with Inequality (7.8) this implies $\left\|f^\star - g^\star\right\|_{\infty;\mathcal{M}} \leq \epsilon'\Lambda \quad +2\sqrt{(CL + \Lambda/2)\,\Lambda\epsilon'}$, conditioned on event $P_{\epsilon'} = \left\{\left\|k - \hat{k}\right\|_\infty \leq \epsilon'\right\}$. For desired $\epsilon > 0$, conditioning on event $P_{\epsilon'}$ with $\epsilon' = \min\left\{\epsilon / \left[2\left(\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda}\right)\right],\ \epsilon^2 / \left[2\left(\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda}\right)\right]^2\right\}$ yields bound $\left\|f^\star - g^\star\right\|_{\infty;\mathcal{M}} \leq \epsilon/2$: if $\epsilon' \leq 1$, then $\epsilon/2 \geq \sqrt{\epsilon'}\left(\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda}\right) \geq \epsilon'\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda\epsilon'}$ provided that $\epsilon' \leq \epsilon^2 / \left[2\left(\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda}\right)\right]^2$. Otherwise if $\epsilon' > 1$, then we have $\epsilon/2 \geq \epsilon'\left(\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda}\right) \geq \epsilon'\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda\epsilon'}$ provided $\epsilon' \leq \epsilon / \left[2\left(\Lambda + 2\sqrt{(CL + \Lambda/2)\,\Lambda}\right)\right]$. Since for any $H > 0$, $\min\left\{H, H^2\right\} \geq \min\left\{1, H^2\right\}$, the result follows. □

We now recall the result due to Rahimi & Recht (2008) that establishes the non-asymptotic uniform convergence of the kernel functions required by the previous lemma (i.e., an upper bound on the probability of event $P_{\epsilon'}$).

LEMMA 7.21 (Rahimi & Recht, 2008, Claim 1) *For any $\epsilon > 0$, $\delta \in (0, 1)$, translation-invariant kernel k, and compact set $\mathcal{M} \subseteq \Re^D$, if $\hat{D} \geq \frac{4(D+2)}{\epsilon^2}\log_e\left(\frac{2^8\left(\sigma_p \mathrm{diam}(\mathcal{M})\right)^2}{\delta\epsilon^2}\right)$, then Algorithm 7.4's random mapping $\hat{\phi}$ from Equation* (7.6) *satisfies* $\Pr\left(\left\|k - \hat{k}\right\|_\infty < \epsilon\right) \geq 1 - \delta$, *where $\sigma_p^2 = \mathrm{E}\left[\langle\boldsymbol{\omega}, \boldsymbol{\omega}\rangle\right]$ is the second moment of the Fourier transform p of k's g function.*

Combining these ingredients establishes utility for PRIVATESVM.

Proof Lemma 7.20 and Corollary 7.18 combined via the triangle inequality with Lemma 7.21, together establish the result as follows. Define $\mathbb{P}$ to be the conditioning event regarding the approximation of k by $\hat{k}$, denote the events in Lemmas 7.20 and

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

7.11 by Q and R, and the target event in the theorem by $\mathbb{S}$.

$$\mathbb{P} = \left\{ \left\| k - \hat{k} \right\|_{\infty;\mathcal{M}} < \min \left\{ 1, \frac{\epsilon^2}{2^2 \left(\Lambda + 2\sqrt{\left(CL + \frac{\Lambda}{2}\right)\Lambda} \right)^2} \right\} \right\}$$

$$\mathbb{Q} = \left\{ \left\| f^\star - \tilde{f} \right\|_{\infty;\mathcal{M}} \leq \frac{\epsilon}{2} \right\}$$
$$\mathbb{R} = \left\{ \left\| \hat{f}^\star - \tilde{f} \right\|_{\infty} \leq \frac{\epsilon}{2} \right\}$$
$$\mathbb{S} = \left\{ \left\| f^\star - \hat{f}^\star \right\|_{\infty;\mathcal{M}} \leq \epsilon \right\}.$$

The claim is a bound on $\Pr(\mathbb{S})$. By the triangle inequality, events $\mathbb{Q}$ and $\mathbb{R}$ together imply $\mathbb{S}$. Second note that event $\mathbb{R}$ is independent of $\mathbb{P}$ and $\mathbb{Q}$. Thus $\Pr(\mathbb{S} \mid \mathbb{P}) \geq \Pr(\mathbb{Q} \cap \mathbb{R} \mid \mathbb{P}) = \Pr(\mathbb{Q} \mid \mathbb{P}) \Pr(\mathbb{R}) \geq 1 \cdot (1 - \delta/2)$, for sufficiently small λ. Finally Lemma 7.21 bounds $\Pr(\mathbb{P})$: Provided that $\hat{D} \geq 4(D+2) \log_e \left(2^9 \left(\sigma_p \text{diam}(\mathcal{M}) \right)^2 / \left(\delta\theta(\epsilon) \right) \right) / \theta(\epsilon)$ where $\theta(\epsilon) = \min \left\{ 1, \epsilon^4 / \left[2 \left(\Lambda + 2\sqrt{(CL + \Lambda/2)\Lambda} \right) \right]^4 \right\}$ we have $\Pr(\mathbb{P}) \geq 1 - \delta/2$. Together this yields $\Pr(\mathbb{S}) = \Pr(\mathbb{S} \mid \mathbb{P}) \Pr(\mathbb{P}) \geq (1 - \delta/2)^2 \geq 1 - \delta$. □

7.7 Bounds on Optimal Differential Privacy

In this section we delve deeper into the special case of the hinge loss. We begin by plugging hinge loss $\ell(y, \hat{y}) = (1 - y\hat{y})_+$ into the main results on privacy and utility of the previous sections. Similar computations can be done for other convex losses; we select hinge loss because it is the most common among SVM classification losses. We then proceed to combine the obtained privacy and utility bounds into an upper bound on the optimal differential privacy for SVM learning with the hinge loss. This notion, while presented here specifically for the SVM, in general quantifies the highest level of privacy achievable over all (ϵ, δ)-useful mechanisms with respect to a target mechanism M.

DEFINITION 7.22 For $\epsilon, C > 0$, $\delta \in (0, 1)$, $N > 1$, loss function $\ell(y, \hat{y})$ convex in $\hat{y}$, and kernel k, the *optimal differential privacy for the SVM* is the function

$$\beta^\star(\epsilon, \delta, C, N, \ell, k) = \inf_{\hat{M} \in \mathcal{I}} \sup_{\left(\mathbb{D}^{(1)}, \mathbb{D}^{(2)}\right) \in \mathcal{D}} \sup_{t \in \mathcal{T}_{\hat{M}}} \log \left(\frac{\Pr\left(\hat{M}\left(\mathbb{D}^{(1)}\right) = t\right)}{\Pr\left(\hat{M}\left(\mathbb{D}^{(2)}\right) = t\right)} \right),$$

where $\mathcal{I}$ is the set of all (ϵ, δ)-useful mechanisms with respect to the SVM with parameter C, loss ℓ, and kernel k; and $\mathcal{D}$ is the set of all pairs of neighboring databases with N entries.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

7.7.1 Upper Bounds

Combining Theorems 7.10 and 7.11 immediately establishes the upper bound on the optimal differential privacy $\beta^\star$ for mechanisms achieving a given desired level (ϵ, δ) of usefulness.

COROLLARY 7.23 *The optimal differential privacy $\beta^\star$ among all mechanisms that are (ϵ, δ)-useful with respect to the* SVM *with finite F-dimensional feature mapping inducing bounded norms $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2$ and $\|\phi(\mathbf{x})\|_\infty \leq \Phi$ for all $\mathbf{x} \in \mathfrak{R}^D$, hinge loss, parameter $C > 0$, on N training, is at most*

$$\begin{aligned}\beta^\star &\leq \frac{8\kappa\Phi C\left(F\log_e 2 + \log_e \frac{1}{\delta}\right)}{N\epsilon} \\ &= \mathcal{O}\left(\frac{C}{\epsilon N}\log\frac{1}{\delta}\right).\end{aligned}$$

Proof The proof is a straightforward calculation for general L-Lipschitz loss. In the general case the bound has the numerator leading coefficient $8\kappa\Phi CL$. The result then follows from the fact that hinge loss is 1-Lipschitz on $\mathfrak{R}$: i.e., $\partial_{\hat{y}}\ell = \mathbf{1}[1 \geq y\hat{y}] \leq 1$. □

Observe that $\Phi \geq \kappa/\sqrt{F}$, so κ could be used in place of Φ to simply the result's statement; however, doing so would yield a slightly looser bound. Also note that by this result, if we set $C = \sqrt{N}$ (needed for universal consistency, see Remark 7.5) and fix β and δ, then the error due to preserving privacy is on the same order as the error in estimating the "true" parameter $\mathbf{w}$.

Recall the dual program for learning under hinge loss from Section 7.3 repeated here for convenience:

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathfrak{R}^N}\ & \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y^{(i)}y^{(j)}k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) \\ \text{s.t. }\ & 0 \leq \alpha_i \leq \frac{C}{N}\ \forall i \in [n].\end{aligned} \tag{7.10}$$

We split the calculation of the upper bound for the translation-invariant kernel case into the following two steps because they are slightly more involved than the finite-dimensional feature mapping case.

COROLLARY 7.24 *Consider any database $\mathbb{D}$ of size N, scalar $C > 0$, and translation-invariant kernel k. For any $\beta > 0$ and $\hat{D} \in \mathfrak{N}$,* PRIVATESVM *run on $\mathbb{D}$ with hinge loss, noise parameter $\lambda \geq \frac{2^{2.5}C\sqrt{\hat{D}}}{\beta N}$, approximation parameter $\hat{D}$, and regularization parameter C guarantees β-differential privacy. Moreover, for any compact set $\mathcal{M} \subseteq \mathfrak{R}^D$ containing $\mathbb{D}$, and scalars $\epsilon > 0$ and $\delta \in (0, 1)$,* PRIVATESVM *run on $\mathbb{D}$ with hinge loss, kernel k, noise parameter $\lambda \leq \min\left\{\frac{\epsilon}{2^4\log_e 2\sqrt{\hat{D}}}, \frac{\epsilon\sqrt{\hat{D}}}{8\log_e\frac{2}{\delta}}\right\}$, approximation parameter $\hat{D} \geq \frac{4(D+2)}{\theta(\epsilon)}\log_e\left(\frac{2^9\left(\sigma_p\mathrm{diam}(\mathcal{M})\right)^2}{\delta\theta(\epsilon)}\right)$ with $\theta(\epsilon) = \min\left\{1, \frac{\epsilon^4}{2^{12}C^4}\right\}$, and parameter C is (ϵ, δ)-useful with respect to hinge-loss* SVM *run on $\mathbb{D}$ with kernel k and parameter C.*

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Proof The first result follows from Theorem 7.10 and the fact that hinge loss is convex and 1-Lipschitz on $\Re$ (as justified in the proof of Corollary 7.23). The second result follows almost immediately from Theorem 7.16. For hinge loss we have that feasible α_i's are bounded by C/N (and so $\Lambda = C$) by the dual's box constraints and that $L = 1$, implying we take $\theta(\epsilon) = \min\left\{1, \frac{\epsilon^4}{2^4 C^4 \left(1+\sqrt{6}\right)^4}\right\}$. This is bounded by the stated $\theta(\epsilon)$. □

Combining the competing requirements on λ upper-bounds optimal differential privacy of hinge-loss SVM.

THEOREM 7.25 *The optimal differential privacy for hinge-loss* SVM *on translation-invariant kernel k is bounded by* $\beta^\star(\epsilon, \delta, C, N, \ell, k) = \mathcal{O}\left(\frac{C}{\epsilon^3 N} \log^{1.5} \frac{C}{\delta\epsilon}\right)$.

Proof Consider hinge loss in Corollary 7.24. Privacy places a lower bound of $\beta \geq 2^{2.5} C\sqrt{\hat{D}}/(\lambda N)$ for any chosen λ, which we can convert to a lower bound on β in terms of ϵ and δ as follows. For small ϵ, we have $\theta(\epsilon) = \mathcal{O}\left(\epsilon^4/C^4\right)$ and so to achieve (ϵ, δ)-usefulness we must take $\hat{D} = \Omega\left(\frac{1}{\epsilon^4} \log_e\left(\frac{C^4}{\delta\epsilon^4}\right)\right)$. There are two cases for utility. The first case is with $\lambda = \epsilon / \left(2^4 \log_e\left(2\sqrt{\hat{D}}\right)\right)$, yielding

$$\begin{aligned}
\beta &= \mathcal{O}\left(\frac{C\sqrt{\hat{D}} \log\sqrt{\hat{D}}}{\epsilon N}\right) \\
&= \mathcal{O}\left(\frac{C}{\epsilon^3 N}\sqrt{\log\frac{C}{\delta\epsilon}}\left(\log\frac{1}{\epsilon} + \log\log\frac{C}{\delta\epsilon}\right)\right) \\
&= \mathcal{O}\left(\frac{C}{\epsilon^3 N} \log^{1.5}\frac{C}{\delta\epsilon}\right).
\end{aligned}$$

In the second case, $\lambda = \frac{\epsilon\sqrt{\hat{D}}}{8 \log_e \frac{2}{\delta}}$ yields $\beta = \mathcal{O}\left(\frac{C}{\epsilon N} \log\frac{1}{\delta}\right)$ which is dominated by the first case as $\epsilon \downarrow 0$. □

A natural question arises from this discussion: given any mechanism that is (ϵ, δ)-useful with respect to hinge SVM, for how small a β can we possibly hope to guarantee β-differential privacy? In other words, what lower bounds exist for the optimal differential privacy for the SVM?

7.7.2 Lower Bounds

Lower bounds peg the level of differential privacy achievable for *any* mechanism approximating SVMs with high accuracy. The following lemma establishes a negative sensitivity result for the SVM mechanism run with the hinge loss and linear kernel.

LEMMA 7.26 *For any $C > 0$ and $N > 1$, there exists a pair of neighboring databases $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$ on N entries, such that the functions $f_1^\star, f_2^\star$ parametrized by SVM run with parameter C, linear kernel, and hinge loss on $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$, respectively, satisfy* $\left\|f_1^\star - f_2^\star\right\|_\infty > \frac{\sqrt{C}}{N}$.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Proof We construct the two databases on the line as follows. Let $0 < m < M$ be scalars to be chosen later. Both databases share negative examples $x_1 = \ldots = x_{\lfloor n/2 \rfloor} = -M$ and positive examples $x_{\lfloor n/2 \rfloor + 1} = \ldots = x_{N-1} = M$. Each database has $x_N = M - m$, with $y^{(N)} = -1$ for $\mathbb{D}^{(1)}$ and $y^{(N)} = 1$ for $\mathbb{D}^{(2)}$. In what follows we use subscripts to denote an example's parent database, so $(x_{i,j}, y^{(i,j)})$ is the j^{th} example from $\mathbb{D}^{(i)}$. Consider the result of running primal SVM on each database:

$$w_1^\star = \underset{w \in \Re}{\operatorname{argmin}} \left[\frac{1}{2} w^2 + \frac{C}{N} \sum_{i=1}^{N} \left(1 - y^{(1,i)} w x_{1,i}\right)_+ \right]$$
$$w_2^\star = \underset{w \in \Re}{\operatorname{argmin}} \left[\frac{1}{2} w^2 + \frac{C}{N} \sum_{i=1}^{N} \left(1 - y^{(2,i)} w x_{2,i}\right)_+ \right].$$

Each optimization is strictly convex and unconstrained, so the optimizing $w_1^\star, w_2^\star$ are characterized by the first-order KKT conditions $0 \in \partial_w f_i(w)$ for f_i being the objective function for learning on $\mathbb{D}^{(i)}$, and ∂_w denoting the subdifferential operator. Now for each $i \in [2]$

$$\partial_w f_i(w) = w - \frac{C}{N} \sum_{j=1}^{N} y^{(i,j)} x_{i,j} \tilde{\mathbf{1}} \left[1 - y^{(i,j)} w x_{i,j}\right],$$

where

$$\tilde{\mathbf{1}}[x] = \begin{cases} \{0\}, & \text{if } x < 0 \\ [0,1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}$$

is the subdifferential of $(x)_+$.

Thus for each $i \in [2]$, we have that $w_i^\star \in \frac{C}{N} \sum_{j=1}^{N} y^{(i,j)} x_{i,j} \tilde{\mathbf{1}} \left[1 - y^{(i,j)} w_i^\star x_{i,j}\right]$ that is equivalent to

$$w_1^\star \in \frac{CM(N-1)}{N} \tilde{\mathbf{1}} \left[\frac{1}{M} - w_1^\star\right] + \frac{C(m-M)}{N} \tilde{\mathbf{1}} \left[w_1^\star - \frac{1}{m-M}\right]$$
$$w_2^\star \in \frac{CM(N-1)}{N} \tilde{\mathbf{1}} \left[\frac{1}{M} - w_2^\star\right] + \frac{C(M-m)}{N} \tilde{\mathbf{1}} \left[\frac{1}{M-m} - w_2^\star\right].$$

The RHSs of these conditions correspond to decreasing piecewise-constant functions, and the conditions are met when the corresponding functions intersect with the diagonal $y = x$ line, as shown in Figure 7.1. If $\frac{C(M(N-2)+m)}{N} < \frac{1}{M}$ then $w_1^\star = \frac{C(M(N-2)+m)}{N}$. And if $\frac{C(MN-m)}{N} < \frac{1}{M}$ then $w_2^\star = \frac{C(MN-m)}{N}$. So provided that

$$\frac{1}{M} > \frac{C(MN-m)}{N} = \max\left\{\frac{C\left(M(N-2)+m\right)}{N}, \frac{C(MN-m)}{N}\right\},$$

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

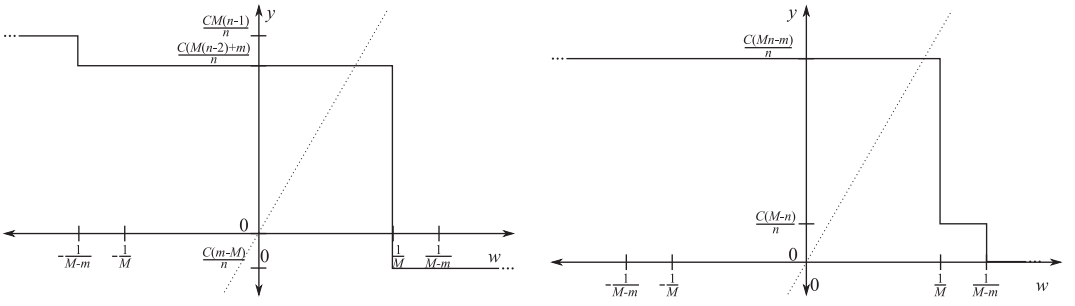

Figure 7.1 For each $i \in [2]$, the SVM's primal solution $w_i^\star$ on database $\mathbb{D}^{(i)}$ constructed in the proof of Lemma 7.26, corresponds to the crossing point of line $y = w$ with $y = w - \partial_w f_i(w)$. Database $\mathbb{D}^{(1)}$ is shown on the left; database $\mathbb{D}^{(2)}$ is shown on the right.

we have $\left|w_1^\star - w_2^\star\right| = \frac{2C}{N}|M - m|$. So taking $M = \frac{2n\epsilon}{C}$ and $m = \frac{N\epsilon}{C}$, this implies

$$\begin{aligned}\left\|f_1^\star - f_2^\star\right\|_\infty &\geq \left|f_{\lfloor}^\star 1\rfloor(1) - f_{\lfloor}^\star 2\rfloor(1)\right| \\ &= \left|w_1^\star - w_2^\star\right| \\ &= 2\epsilon,\end{aligned}$$

provided $\epsilon < \frac{\sqrt{C}}{2N}$. In particular taking $\epsilon = \frac{\sqrt{C}}{2N}$ yields the result. □

With this negative sensitivity result in hand, we can lower bound the optimal differential privacy for any mechanism approximating the SVM with hinge loss.

THEOREM 7.27 (Lower bound on optimal differential privacy for linear SVM) *For any $C > 0$, $N > 1$, $\delta \in (0, 1)$, and $\epsilon \in \left(0, \frac{\sqrt{C}}{2N}\right)$, the optimal differential privacy for the hinge-loss* SVM *with linear kernel is lower bounded by* $\log_e \frac{1-\delta}{\delta} = \Omega\left(\log \frac{1}{\delta}\right)$.

Proof Consider (ϵ, δ)-useful mechanism $\hat{M}$ with respect to SVM learning mechanism M with parameter $C > 0$, hinge loss, and linear kernel on N training examples, where $\delta > 0$ and $\frac{\sqrt{C}}{2N} > \epsilon > 0$. By Lemma 7.26 there exists a pair of neighboring databases $\mathbb{D}^{(1)}, \mathbb{D}^{(2)}$ on N entries, such that $\left\|f_1^\star - f_2^\star\right\|_\infty > 2\epsilon$ where $f_i^\star = f_{M\left(\mathbb{D}^{(i)}\right)}$ for each $i \in [2]$. Let $\hat{f}_i = f_{\hat{M}\left(\mathbb{D}^{(i)}\right)}$ for each $i \in [2]$. Then by the utility of $\hat{M}$,

$$\Pr\left(\hat{f}_1 \in \mathcal{B}_\epsilon^\infty\left(f_1^\star\right)\right) \geq 1 - \delta, \tag{7.11}$$

$$\Pr\left(\hat{f}_2 \in \mathcal{B}_\epsilon^\infty\left(f_1^\star\right)\right) \leq \Pr\left(\hat{f}_2 \notin \mathcal{B}_\epsilon^\infty\left(f_2^\star\right)\right) < \delta. \tag{7.12}$$

Let $\hat{\mathcal{P}}_1$ and $\hat{\mathcal{P}}_2$ be the distributions of $\hat{M}\left(\mathbb{D}^{(1)}\right)$ and $\hat{M}\left(\mathbb{D}^{(2)}\right)$, respectively, so that $\hat{\mathcal{P}}_i(t) = \Pr\left(\hat{M}\left(\mathbb{D}^{(i)}\right) = t\right)$. Then by Inequalities (7.11) and (7.12)

$$\mathrm{E}_{T\sim\mathcal{P}_1}\left[\frac{d\mathcal{P}_2(T)}{d\mathcal{P}_1(T)} \,\middle|\, T \in \mathcal{B}_\epsilon^\infty\left(f_1^\star\right)\right] = \frac{\int_{\mathcal{B}_\epsilon^\infty(f_1^\star)} \frac{d\mathcal{P}_2(t)}{d\mathcal{P}_1(t)} d\mathcal{P}_1(t)}{\int_{\mathcal{B}_\epsilon^\infty(f_1^\star)} d\mathcal{P}_1(t)} \leq \frac{\delta}{1-\delta}.$$

Thus there exists a t such that $\log \frac{\Pr\left(\hat{M}\left(\mathbb{D}^{(1)}\right)=t\right)}{\Pr\left(\hat{M}\left(\mathbb{D}^{(2)}\right)=t\right)} \geq \log \frac{1-\delta}{\delta} = \Omega\left(\log \frac{1}{\delta}\right)$. □

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Remark 7.28 Equivalently this result can be written as follows. For any $C > 0$, $\beta > 0$, and $N > 1$, if a mechanism $\hat{M}$ is (ϵ, δ)-useful and β-differentially private, then either $\epsilon \geq \frac{\sqrt{C}}{2N}$ or $\delta \geq \exp(-\beta)$.

We have now presented both upper bounds (Corollary 7.23 with $L = 1$) and lower bounds on the optimal differential privacy for the case of the linear SVM with hinge loss. Ignoring constants and using the scaling of C (see Remark 7.5) we have that

$$\Omega\left(\log \frac{1}{\delta}\right) = \beta^{\star} = \mathcal{O}\left(\frac{1}{\epsilon\sqrt{N}} \log \frac{1}{\delta}\right).$$

It is noteworthy that the bounds agree in their scaling on utility confidence δ, but that they disagree on linear and square-root terms in their dependence on ϵ and N, respectively. Moreover under the appropriate scaling of C, the lower bound holds only for $\epsilon = \mathcal{O}\left(N^{-0.75}\right)$; under which the upper asymptotic bound becomes $\mathcal{O}\left(N^{0.25} \log(1/\delta)\right)$. Finding better-matching bounds remains an interesting open problem.

We refer the reader to Rubinstein et al. (2012) for a similar lower bound under the RBF kernel. There a negative sensitivity result is achieved not through a pair of neighboring databases that induce very different SVM results, but through a sequence of K pairwise-neighboring databases whose images under SVM learning form an ϵ-packing.

7.8 Summary

In this chapter we presented mechanisms for private SVM learning due to Chaudhuri et al. (2011) and Rubinstein et al. (2012), which release a classifier based on a privacy-sensitive database of training data. The former approach is one of objective perturbation while the latter performs output perturbation, calibrated by the algorithmic stability of regularized ERM—a property that is typically used in learning theory to prove risk bounds of learning algorithms.

In addition to measuring the training data differential privacy preserved by the output perturbation mechanisms, we also focused on their utility: the similarity of the classifiers released by private and nonprivate SVM. This form of utility implies good generalization error of the private SVM. To achieve utility under infinite-dimensional feature mappings, both families of approach perform regularized empirical risk minimization (ERM) in a random reproducing kernel Hilbert space whose kernel approximates the target kernel. This trick, borrowed from large-scale learning, permits the mechanisms to privately respond with a finite representation of a maximum-margin hyperplane classifier. We explored the high-probability, pointwise similarity between the resulting function and the nonprivate SVM classifier through a smoothness result of regularized ERM with respect to perturbations of the RKHS.

Interesting directions involve extending the ideas of this chapter to other learning algorithms:

Question 7.1 Can the mechanisms and proof techniques used for differentially private SVM by output perturbation be extended to other kernel methods?

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press

Question 7.2 Is there a general connection between algorithmic stability and global sensitivity?

Such a connection would immediately suggest a number of practical privacy-preserving learning mechanisms for which calculations on stability are available: stability would dictate the level of (possibly Laplace) noise required for differential privacy, and for finite-dimensional feature spaces utility would likely follow a similar pattern as presented here for the SVM. The application of the random RKHS with kernel approximating a target kernel would also be a useful tool in making kernelized learners differentially private for translation-invariant kernels.

Bounds on differential privacy and utility combine to upper bound the optimal level of differential privacy possible among all mechanisms that are (ϵ, δ)-useful with respect to the hinge-loss SVM. Lower bounds on this quantity establish that any mechanism that is too accurate with respect to the hinge SVM, with any nontrivial probability, cannot be β-differentially private for small β.

Question 7.3 An important open problem is to reduce the gap between upper and lower bounds on the optimal differential privacy of the SVM.

https://doi.org/10.1017/9781107338548.007 Published online by Cambridge University Press