

# Contraction of $E_\gamma$ -Divergence and Its Applications to Privacy

Shahab Asoodeh, Mario Diaz, and Flavio P. Calmon

## Abstract

We investigate the contraction coefficients derived from strong data processing inequalities for the  $E_\gamma$ -divergence. By generalizing the celebrated Dobrushin's coefficient from total variation distance to  $E_\gamma$ -divergence, we derive a closed-form expression for the contraction of  $E_\gamma$ -divergence. This result has fundamental consequences in two privacy settings. First, it implies that local differential privacy can be equivalently expressed in terms of the contraction of  $E_\gamma$ -divergence. This equivalent formula can be used to precisely quantify the impact of local privacy in (Bayesian and minimax) estimation and hypothesis testing problems in terms of the reduction of effective sample size. Second, it leads to a new information-theoretic technique for analyzing privacy guarantees of online algorithms. In this technique, we view such algorithms as a composition of amplitude-constrained Gaussian channels and then relate their contraction coefficients under  $E_\gamma$ -divergence to the overall differential privacy guarantees. As an example, we apply our technique to derive the differential privacy parameters of gradient descent. Moreover, we also show that this framework can be tailored to batch learning algorithms that can be implemented with one pass over the training dataset.

## I. INTRODUCTION

Consider the Markov chain

$$X_1 \rightarrow Y_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \rightarrow Y_n \rightarrow X_{n+1}, \quad (1)$$

where the  $d$ -dimensional random variable  $X_1$  is the input of the first Markov kernel (or channel)  $P_{Y_1|X_1}$  and is assumed to satisfy  $\|X_1\|^2 \leq dA$  almost surely (a.s.) for some  $A > 0$ . Each kernel  $P_{X_{t+1}|X_t}$  is composed of the concatenation of two channels:

- 1)  $P_{Y_t|X_t}$  is a vector-Gaussian kernel of dimension  $d$ , i.e.,

$$P_{Y_t|X_t=x} = \mathcal{N}(x, \sigma_t^2 \mathbf{I}_d), \quad (2)$$

- 2)  $P_{X_{t+1}|Y_t}$  is an arbitrary kernel that ensures  $X_{t+1}$  satisfies the same amplitude constraint as  $X_1$ . Thus, we have for all  $t \in [n+1] := \{1, \dots, n+1\}$

$$\|X_t\|^2 \leq dA, \quad \text{a.s.} \quad (3)$$

For example,  $P_{X_{t+1}|Y_t}$  can be the projection of  $Y_t$  onto the  $\ell_2$ -ball of radius  $(dA)^{1/2}$ .

One specific instantiation of this Markov chain will be given in Section IV (see Fig. 2) for characterizing the privacy guarantee of the online gradient descent algorithm, whose each iterations is modelled by the composition of a Gaussian kernel and a projection operator.

Let  $\mu_{n+1}$  and  $\mu'_{n+1}$  be the distributions of the output  $X_{n+1}$  of chain (1) when  $X_1$  has distribution  $\mu_1$  and  $\mu'_1$ , respectively. The main goal of this paper is to characterize the divergence between  $\mu_{n+1}$  and

This work was supported in part by NSF under grants CIF 1900750 and CIF CAREER 1845852. Parts of the results in this paper were presented at the International Symposium on Information Theory 2020 and 2021 [1, 2].

S. Asoodeh and F. P. Calmon are with School of Engineering and Applied Science, Harvard University (e-mails: {shahab, flavio}@seas.harvard.edu). M. Diaz is with the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, Mexico City 04510, Mexico (e-mail: mario.diaz@sigma.iimas.unam.mx).

$\mu'_{n+1}$  in terms of  $\mu_1$  and  $\mu'_1$ . More specifically, we derive *strong data processing inequalities* [3] for Gaussian channels satisfying the amplitude constraint (3).

When measuring the distance between  $\mu_{n+1}$  and  $\mu'_{n+1}$  via  $f$ -divergences, this framework has been extensively studied in information theory literature [3–10]. Given a convex function  $f : (0, \infty) \rightarrow \mathbb{R}$  such that  $f(1) = 0$ , the  $f$ -divergence between two probability measures  $\mu$  and  $\nu$  with  $\mu \ll \nu$  is defined in [11, 12] as

$$D_f(\mu\|\nu) := \mathbb{E}_\nu \left[ f \left( \frac{d\mu}{d\nu} \right) \right].$$

Specific instances of  $f$ -divergences include KL-divergence, where  $f(t) = t \log(t)$ , and total variation, in which case  $f(t) = \frac{1}{2}|t - 1|$ . Let  $\mathcal{D}$  and  $\mathcal{W}$  be subsets of (potentially different) Euclidean spaces and let  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  be a Markov kernel (i.e., channel), where  $\mathcal{P}(\mathcal{W})$  denotes the set of all probability measures on  $\mathcal{W}$ . Following the information-theoretic approach described in Ahlswede and Gács [3], we define the *contraction coefficient* (or strong data processing coefficient) of  $K$  under  $D_f(\cdot\|\cdot)$  as

$$\eta_f(K) := \sup_{\substack{\mu, \nu \in \mathcal{P}(\mathcal{D}): \\ D_f(\mu\|\nu) \neq 0}} \frac{D_f(\mu K\|\nu K)}{D_f(\mu\|\nu)}, \quad (4)$$

where  $\mu K$  and  $\nu K$  denote the distribution on  $\mathcal{W}$  induced by the push-forward of  $\mu$  and  $\nu$ , respectively. This quantity has been extensively studied for general channels in [3–7] and for Gaussian channels with power and amplitude constraints (see e.g., [8–10]). Most notably, Dobrushin [4] showed that  $\eta_{TV}$ —the contraction coefficient under total variation distance—has a remarkably simple expression:

$$\eta_{TV}(K) = \sup_{x_1, x_2 \in \mathcal{D}} TV(K(\cdot|x_1), K(\cdot|x_2)). \quad (5)$$

This formula has found several applications in the study of ergodicity of Markov processes as well as Gibbs measures, e.g., see [4, 5, 7, 8].

Throughout this work, we focus on an instantiation of  $f$ -divergence named  $E_\gamma$ -divergence (also known as hockey-stick divergence) [13–15]. Given  $\gamma \geq 0$ , the  $E_\gamma$ -divergence between two probability measures  $\mu$  and  $\nu$  in  $\mathcal{P}(\mathcal{D})$  is defined as

$$E_\gamma(\mu\|\nu) := \int d(\mu - \gamma\nu)^+ - (1 - \gamma)^+ \quad (6)$$

$$= \sup_{A \subset \mathcal{D}} [\mu(A) - \gamma\nu(A)] - (1 - \gamma)^+ \quad (7)$$

$$= \mu(\iota_{\mu\|\nu} > \log \gamma) - \gamma\nu(\iota_{\mu\|\nu} > \log \gamma) - (1 - \gamma)^+, \quad (8)$$

where  $(\mu - \gamma\nu)^+$  is the positive part of the signed measure  $\mu - \gamma\nu$  and  $\iota_{\mu\|\nu}(t) := \log \frac{d\mu}{d\nu}(t)$  denotes the *information density* between  $\mu$  and  $\nu$ . It can be directly verified that  $E_\gamma$ -divergence is in fact the  $f$ -divergence associated with

$$f(t) = (t - \gamma)^+ - (1 - \gamma)^+,$$

where  $(a)^+ := \max(0, a)$ , and also that

$$E_1(\mu\|\nu) = TV(\mu, \nu).$$

We adopt  $E_\gamma$ -divergence for two main reasons:

- Since it is a generalization of total variation distance, its contraction coefficient can potentially broaden the applicability of Dobrushin's result (5),
- As we will see in Sections III and IV,  $E_\gamma$ -divergence has a close connection with both local and central differential privacy (see Theorems 3 and Definition 4). A characterization of the contraction coefficient of  $E_\gamma$ -divergence leads to a simple and precise privacy analysis of several applications in statistics and machine learning.

Our main result (i.e., Theorem 2) is a formula for the contraction coefficient of channels under  $E_\gamma$ -divergence which extends Dobrushin's result (5) from total variation distance to  $E_\gamma$ -divergence for any  $\gamma \geq 0$ . This result has several direct consequences. First, it allows us to write an equivalent expression for local differential privacy, which lends itself well to studying the (minimax and Bayesian) estimation and testing problems under local differential privacy. Following this path, we quantify the impact of local privacy in such problems in terms of the reduction of effective sample size. Second, it enables us to derive a sharp upper bound for the deviation of  $\mu_{n+1}$  from  $\mu'_{n+1}$ , i.e.,  $E_\gamma(\mu_{n+1} \| \mu'_{n+1})$ , in terms of the amplitude constraint  $A$  and each channel's noise variance — a result which turns out instrumental in the privacy analysis of iterative algorithms in Section IV.

**Theorem 1.** *Let  $X_1 \sim \mu_1$  and  $X'_1 \sim \mu'_1$  be two inputs of the Markov chain (1) where channels  $\{P_{X_{t+1}|X_t}\}_{t=1}^n$  satisfy (2) and (3). Let  $X_{n+1} \sim \mu_{n+1}$  and  $X'_{n+1} \sim \mu'_{n+1}$  be the output of the Markov chain when the input is  $X_1$  and  $X'_1$ , respectively. Then, we have*

$$E_\gamma(\mu_{n+1} \| \mu'_{n+1}) \leq E_\gamma(\mu_1 \| \mu'_1) \prod_{t=1}^n \theta_{\gamma \vee \frac{1}{\gamma}}\left(\frac{2dA}{\sigma_t}\right),$$

where  $\theta_\gamma(r) := Q\left(\frac{\log \gamma}{r} - \frac{r}{2}\right) - \gamma Q\left(\frac{\log \gamma}{r} + \frac{r}{2}\right)$  and  $Q(a) := \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du$ .

When  $\gamma = 1$  and  $\sigma_t = \sigma$ , this theorem reduces to the estimate

$$\text{TV}(\mu_{n+1}, \mu'_{n+1}) \leq \text{TV}(\mu_1, \mu'_1)(1 - 2Q(\sqrt{dA}))^n,$$

that was derived in [8] by a direct application of Dobrushin's result (5). However, unlike [8], we are interested in settings where  $\gamma$  need not be equal to 1. In particular, we make use of Theorem 1 to study the cost of differential privacy in online algorithms such as online gradient descent in which  $\gamma$  is set to be  $e^\varepsilon$ , where  $\varepsilon \geq 0$  is the differential privacy guarantee.

### A. Main Contributions

**Contraction coefficient under  $E_\gamma$ -divergence:** We prove that, similar to  $\eta_{\text{TV}}$ , the contraction coefficient of a Markov kernel  $K$  under  $E_\gamma$ -divergence, denoted by  $\eta_\gamma(K)$ , enjoys a remarkably simple two-point characterization. More precisely, we show in Theorem 2 that

$$\eta_\gamma(K) = \sup_{x_1, x_2 \in \mathcal{D}} \mathbb{E}_\gamma(K(\cdot|x_1) \| K(\cdot|x_2)),$$

for all  $\gamma \geq 1$ , thus generalizing (5) from total variation distance (i.e.,  $\gamma = 1$ ) to  $E_\gamma$ -divergence for any  $\gamma > 1$ . We then use basic properties of  $E_\gamma$ -divergence to show  $\eta_\gamma(K) = \eta_{1/\gamma}(K)$  for  $\gamma < 1$ .

We apply this result to the integral representation of  $f$ -divergences in terms of  $E_\gamma$ -divergence [16, Corollary 3.7] to obtain the estimate

$$D_f(\mu K \| \nu K) \leq \int_0^\infty \eta_\gamma(K) f''(\gamma) E_\gamma(\mu \| \nu) d\gamma, \quad (9)$$

for all  $\mu, \nu \in \mathcal{P}(\mathcal{D})$  and all  $f$ -divergences with twice-differentiable  $f$ . This result is provably tighter than the bound [16, Proposition II.4.10]

$$D_f(\mu K \| \nu K) \leq \eta_{\text{TV}}(K) D_f(\mu \| \nu).$$

While this upper bound is known to be typically strict, it has been used extensively in the information theory literature [8, 17, 18]. We numerically compare these two upper bounds in Examples 1 and 2 for  $\chi^2$ -divergence.

**Local differential privacy:** As another application, we use Theorem 2 to show that local differential privacy (LDP) can be equivalently expressed in terms of the contraction coefficient under  $E_\gamma$ -divergence (see Theorem 3): A randomized mechanism  $K$  is  $(\varepsilon, \delta)$ -LDP if and only if

$$\eta_\gamma(K) \leq \delta \text{ for } \gamma = e^\varepsilon.$$

Using the relationship between  $E_\gamma$ -divergence and general  $f$ -divergences, this result implies

$$\eta_f(K) \leq 1 - (1 - \delta)e^{-\varepsilon}$$

for any  $(\varepsilon, \delta)$ -LDP mechanism  $K$ . This estimate of the output  $f$ -divergence of an LDP mechanism can be directly used to quantify the impact of local privacy in Bayesian and minimax estimation problems. First, we introduce a new variant of Le Cam's converse technique [19] to derive a lower bound for the minimax estimation risk under LDP constraints. In contrast to existing results on minimax risk under LDP [20–25], our result holds for any values of  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ .

Second, we develop a framework for characterizing the Bayesian estimation risk under LDP. To do so, we first derive a lower bound for the non-private Bayesian risk in terms of  $E_\gamma$ -information, defined as the  $E_\gamma$ -divergence between the joint and product distributions of two random variables (i.e., the  $E_\gamma$  counterpart of mutual information). Combining this bound with Theorem 2, we derive a lower bound for private Bayesian risk in Theorem 4. Our results indicate that the cost of  $(\varepsilon, \delta)$ -LDP in one-dimensional minimax and Bayesian estimation problems is to reduce the effective sample size from  $n$  to  $n(1 - (1 - \delta)e^{-\varepsilon})$ .

**Central differential privacy:** As a final application, we apply Theorem 2 (more specifically, Theorem 1) to investigate the privacy guarantees of online iterative algorithms. In general, an online learning algorithm proceeds as follows: a learner first selects a random point  $W_1$  from a convex set  $\mathcal{W} \subset \mathbb{R}^d$ . After committing to  $W_1$ , the cost function  $f_1$  is revealed to her by nature, specifying the cost  $f_1(W_1)$  incurred by the choice  $W_1$ . Upon observing the cost function  $f_t$  at time  $t$ , the learner constructs  $W_{t+1} \in \mathcal{W}$  at time  $t + 1$  according to some update rules based *only* on  $f_t$  (as opposed to the entire  $\{f_1, \dots, f_t\}$ ). Denoting this update rule at time  $t$  by  $\Psi_{f_t} : \mathcal{W} \rightarrow \mathbb{R}^d$ , the iterative algorithm can be expressed by

$$W_{t+1} = \Pi_{\mathcal{W}}(\Psi_{f_t}(W_t)),$$

where  $\Pi_{\mathcal{W}}(\cdot)$  is the projection operator onto  $\mathcal{W}$  (see Section IV for examples of this algorithms). To ensure privacy of the learner, one standard way is to add calibrated noise to  $\Psi_{f_t}$  at each iteration [26–29]. Thus, the private version of such algorithm can typically be expressed as

$$W_{t+1} = \Pi_{\mathcal{W}}(\Psi_{f_t}(W_t) + \sigma_t Z_t), \quad (10)$$

where  $\{Z_t\}$  is the collection of independent and identically distributed (i.i.d.) noise variables sampled from a known density with covariance matrix  $I_d$  and  $\sigma_t$  specifies the magnitude of noise at time  $t$ . In Theorem 6, we characterize the central differential privacy (DP) guarantee of such algorithm when the learner discloses  $W_{n+1}$  and  $\{Z_t\}$  are sampled from the Gaussian distribution. Viewing each iteration of this algorithm as a composition of a Gaussian kernel and a projection operator (see (2) and (3)), we precisely model this iterative algorithm by the Markov chain (1). Invoking Theorem 1, we can thus obtain an upper bound on the  $E_\gamma$ -divergence between the distributions of the output of the process (10) after  $n$  iterations when cost function  $f_j$  changes to  $f'_j$  for some  $j \in [n]$ . The relationship between  $E_\gamma$ -divergence and DP enables us to directly translate this bound to a bound on the DP parameters  $\varepsilon$  and  $\delta$ . We instantiate Theorem 6 to characterize privacy guarantees in *one-pass* stochastic gradient descent (SGD) in Corollary 5 and the online gradient descent in Proposition 2.

## B. Additional Related Work

Strong data processing inequalities (SDPI) for KL divergence and total variation distance are ubiquitous in information theory and statistics. They appear, for example, in the study of ergodicity of Markov

processes [4], the uniqueness of Gibbs measures [4], contraction of mutual information (and generalized mutual information) in Markov chains [5, 8, 10, 30, 31] and in Bayesian networks [32], comparison of channels [33], distributed estimation [34], communication complexity of statistical estimation [35], distributed function computation [36], and private estimation problems [22, 37]. A formula for the contraction coefficients under general  $f$ -divergences is derived in [7, Theorem 5.2] for differentiable  $f$ . This formula, however, is not applicable for the specific case of  $E_\gamma$ -divergence as  $f(t) = (t - \gamma)^+$  is not differentiable. More recently, Kamalaruban [38, Theorem 3.10] derived a closed-form expression for the contraction coefficient under another  $f$ -divergence, namely *DeGroot's statistical information* [39].

The study of statistical efficiency under *pure* LDP constraints was initiated by Duchi et al. [20] in the minimax setting and has since gained considerable attention, e.g., [21–25, 40–48]. While the original bounds on the private minimax risk in [20] were meaningful only in the high privacy regime (i.e., small  $\epsilon$ ), the order optimal bounds were recently given by Duchi and Rogers in [22] for the general privacy regime. Interestingly, their technique relies on the decay rate of mutual information over a Markov chain, which is known to be equivalent to the SDPI for KL divergence [30]. Unlike their technique, ours is based on the SDPI for  $E_\gamma$ -divergence and allows us to study minimax risk under *approximate* LDP (i.e.,  $\delta > 0$ ).

Online learning [49, 50] is a framework where a sequence of predictions are made given the knowledge of past actions. This framework is particularly well-suited for dynamic and adversarial environments where learning from data must be done in real-time. Online learning is ubiquitous in practical applications such as recommender systems [51], spam detection [52], portfolio optimization [53–55], and convex optimization [56, 57], to name a few. The problem of analyzing DP guarantees of online algorithms was introduced in [58] for the simple setting where  $\mathcal{W}$  is a probability simplex and the cost functions are linear with binary coefficients. Inspired by this work, [26–28] developed techniques for making a large class of online learning algorithms differentially private. The settings studied in these works differs from ours in that they violate two implicit assumptions made in our setting:

- 1) Each update  $W_{t+1}$  depends only on  $f_t$ , conditioned on  $W_t$ . However, the aforementioned works rely on the tree-based aggregation [58] which selects  $W_{t+1}$  based on the entire  $f_1, \dots, f_t$ .
- 2) The output of each iteration is kept private. Thus the privacy guarantee must be quantified against  $W_{n+1}$  and not the entire  $W_1, \dots, W_{n+1}$ .

The Markovity condition in Assumption 1 is equivalent to a memory constraint on the online algorithm. Moreover, it is trivially satisfied by several popular online algorithms such as online (stochastic) gradient descent [59], online mirror descent [49], implicit gradient descent [60], the passive-aggressive algorithm [61], composite mirror descent [62], and the Frank-Wolfe algorithm [63]. Nevertheless, this assumption rules out algorithms such as the follow-the-leader (FTL) algorithm and its variants. We do note, however, that a particular popular variant of the FTL algorithm, namely Regularized FTL [64, 65] is equivalent to online mirror descent (see [66, Lemma 1] and [67]), rendering our framework applicable in this specific case.

Assumption 2 is also present in several recent works on last-iterate convergence of learning algorithms [68–72], the *privacy amplification by iteration* framework [73], its variants in [74, 75], and privacy-preserving generative model for data inspection [76]. Another practical scenario where Assumption 2 naturally arises is as follows: A learner intends to fit a model privately and publicly releases the model parameters after a target level of accuracy is met (for instance, after a certain number of iterations). Hiding intermediate updates not only leads to privacy amplification [73, 75], it may also drive the final parameters to be sparse [72] — a property which is often crucial for many applications [72].

The iterative process (10) can also correspond to batch (i.e., offline) algorithms where a dataset is fixed and is available to the learner in advance. For instance, consider the empirical risk minimization (ERM) problem: Given a dataset  $\{x_1, \dots, x_n\} \in \mathcal{X}^n$ , a learner seeks to solve  $\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{t=1}^n \ell(w, x_t)$  for some convex loss function  $\ell : \mathcal{W} \times \mathcal{X} \rightarrow \mathbb{R}_+$ . The problem of ERM with DP constraints has been widely-

studied with known asymptotically tight upper and lower bounds for the excess loss (i.e., the difference between achieved loss and the true minimum), e.g., [77–88]. This problem can be viewed as an instance of the online learning problem with cost functions  $f_t(w) = \ell(w, x_t)$  [89]. This observation enables us to translate the privacy analysis of the process (10) to the privacy guarantee of batch algorithms. The caveat here is that each data point  $x_i, i \in [n]$  must be involved in the training process exactly once. Examples of such algorithms include SGD with sub-sampling *without* replacement.

The privacy properties of SGD algorithms with Gaussian perturbations under similar assumptions as ours was initiated in [73] where the privacy guarantee was given in terms of a variant of DP, namely *Rényi differential privacy* (RDP) [90]. The RDP guarantee can easily be converted to the DP guarantee. Recently, an *optimal* conversion formula between RDP and DP was presented in [91]. In Section IV-B, we revisit the model studied in [73], and analytically and numerically demonstrate that our result can be substantially tighter than what would be obtained by converting their results from RDP to DP via the formula given in [91].

### C. Paper Organization

The rest of the paper is organized as follows. Section II presents our main result on the contraction coefficient  $\eta_\gamma$ , its multi-letter generalizations, and an explicit formula for contraction in amplitude-constrained additive Gaussian kernels. Section II shows how to equivalently express LDP constraints in terms of  $\eta_\gamma$ . This section also presents three applications of  $E_\gamma$ -contraction to quantify the impact of LDP in estimation in terms of minimax risk (Section III-A), Bayesian risk (Section III-B), and binary hypothesis testing (Section III-C). Section IV demonstrates how typical machine-learning iterative algorithms can be viewed as a composition of Markov kernels and how their contraction coefficients can be used to obtain tighter privacy guarantees for projected noisy stochastic gradient descent in Section IV-B and online gradient descent in Section IV-C.

### D. Notation

We use upper-case letters (e.g.,  $X$ ) to denote random variables and calligraphic letters to represent their alphabets (e.g.,  $\mathcal{X}$ ). The set of all distributions on  $\mathcal{X}$  is denoted by  $\mathcal{P}(\mathcal{X})$ . For a signed measure  $\phi$  over  $\mathcal{X}$ , its total variation is defined as

$$\|\phi\| := \phi^+(\mathcal{X}) + \phi^-(\mathcal{X}),$$

where  $(\phi^+, \phi^-)$  is the Hahn-Jordan decomposition of  $\phi$  (see, e.g., [92, Page 421]). Observe that if  $\mu$  and  $\nu$  are probability measures,

$$\|\mu - \nu\| = 2 \sup_{A \subset \mathcal{X}} [\mu(A) - \nu(A)].$$

The total variation distance between two distributions  $\mu$  and  $\nu$  is

$$\text{TV}(\mu, \nu) := \frac{1}{2} \|\mu - \nu\|.$$

A Markov kernel (or channel)  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  is specified by a collection of distributions  $\{K(x) \in \mathcal{P}(\mathcal{W}) : x \in \mathcal{S}\}$ . Given a Markov kernel  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  and  $\mu \in \mathcal{P}(\mathcal{D})$ , we denote by  $\mu K$  the push-forward of  $\mu$  under  $K$ , i.e., the output distribution of  $K$  when the input is distributed according to  $\mu$ , and is given by  $\mu K := \int \mu(dx) K(x)$ . We use  $\mathbb{E}_\mu[\cdot]$  to write the expectation with respect to  $\mu$  and  $[n]$  for an integer  $n \geq 1$  to denote  $\{1, \dots, n\}$ . For  $a \in [0, 1]$ , we define  $\bar{a} = 1 - a$ . For a set  $\mathcal{D} \subset \mathbb{R}^d$ , we let  $\text{dia}(\mathcal{D})$  be its diameter, i.e.,

$$\text{dia}(\mathcal{D}) = \sup_{x_1, x_2 \in \mathcal{D}} \|x_2 - x_1\|.$$

## II. CONTRACTION COEFFICIENT OF MARKOV KERNELS UNDER $E_\gamma$ -DIVERGENCE

In this section, we establish a closed-form expression for the contraction coefficient of Markov kernels under  $E_\gamma$ -divergence which generalizes Dobrushin's formula. Relying on this expression, we compute the contraction coefficient of the amplitude-constrained additive Gaussian kernel. Finally, we improve classical estimates for the output  $f$ -divergence based on our generalization of Dobrushin's formula.

Similar to Dobrushin's formula (5), the next theorem reduces the computation of the contraction coefficient of a Markov kernel under  $E_\gamma$ -divergence to maximizing the output  $E_\gamma$ -divergence for atomic inputs. In the sequel, we always assume that  $\mathcal{D}$  and  $\mathcal{W}$  are subsets of (potentially different) Euclidean spaces. Recall that  $\eta_\gamma(K)$  denotes the contraction coefficient of kernel  $K$  under  $E_\gamma$ -divergence.

**Theorem 2.** *Let  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  be a Markov kernel. For any  $\gamma \geq 1$ , we have*

$$\eta_\gamma(K) = \sup_{x_1, x_2 \in \mathcal{D}} E_\gamma(K(\cdot|x_1) \| K(\cdot|x_2)). \quad (11)$$

Furthermore, for any  $\gamma < 1$ , we have

$$\eta_\gamma(K) = \eta_{1/\gamma}(K). \quad (12)$$

The proof of this theorem is given in Appendix A. In view of the reciprocity relation (12), in the sequel we focus on the case  $\gamma \geq 1$ .

Liu et al. [93, Proposition 4] showed that, for  $\gamma \geq 1$  and any probability measures  $\mu$  and  $\nu$ ,

$$TV(\mu, \nu) \leq 1 - \frac{1 - E_\gamma(\mu \| \nu)}{\gamma}.$$

Applying Theorem 2 to this inequality leads to the following general relation between the contraction coefficient of a Markov kernel under total variation and the corresponding one under  $E_\gamma$ -divergence.

**Lemma 1.** *Let  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  be a Markov kernel. For any  $\gamma \geq 1$ , we have*

$$\eta_{TV}(K) \leq 1 - \frac{1 - \eta_\gamma(K)}{\gamma}.$$

Recall that, for any convex function  $f$  satisfying  $f(1) = 0$  and any Markov kernel  $K$ ,

$$\eta_f(K) \leq \eta_{TV}(K). \quad (13)$$

This result, originally proved in [16] for the discrete case and subsequently extended to the general case in [94], and our previous lemma produce the estimate

$$\eta_f(K) \leq 1 - \frac{1 - \eta_\gamma(K)}{\gamma},$$

or, equivalently, for any probability measures  $\mu$  and  $\nu$ ,

$$D_f(\mu K \| \nu K) \leq \left(1 - \frac{1 - \eta_\gamma(K)}{\gamma}\right) D_f(\mu \| \nu). \quad (14)$$

In Section III we apply the latter bound to quantify the statistical cost of local differential privacy.

Next we extend Lemma 1 to the tensor product of Markov kernels. Given Markov kernels  $K_1, \dots, K_n$ , we let  $K^{\otimes n}$  be their  $n$ -fold tensor product, i.e., the memoryless channel defined by

$$K^{\otimes n} = K_1 \otimes \cdots \otimes K_n.$$

Recall that the contraction coefficient  $\eta_{TV}$  satisfies (see, e.g., [32, Corollary 9], [18, Lemma 3] and [6, Eq. (62)]),

$$\eta_{TV}(K^{\otimes n}) \leq \max_{i \in [n]} [1 - (1 - \eta_{TV}(K_i))^n].$$

The next lemma is an immediate consequence of this inequality and Lemma 1.

**Lemma 2.** *Let  $K_1, \dots, K_n$  be Markov kernels. For any  $\gamma \geq 1$ , we have*

$$\eta_{\text{TV}}(K^{\otimes n}) \leq \max_{i \in [n]} \left[ 1 - \left( \frac{1 - \eta_\gamma(K_i)}{\gamma} \right)^n \right].$$

As before, we can combine the previous lemma with (13) to obtain an upper bound for  $\eta_f(K^{\otimes n})$  with  $f$  a convex function satisfying  $f(1) = 0$ .

#### A. Contraction Coefficient of the Amplitude-Constrained Additive Gaussian Kernel

Next we compute the contraction coefficient of the amplitude-constrained additive Gaussian kernel under  $E_\gamma$ -divergence. To this end, we start by recalling the following lemma.

**Lemma 3** ([95, Lemma 6]). *For  $m \in \mathbb{R}^d$  and  $\sigma > 0$ , let  $\mathcal{N}(m, \sigma^2 \mathbf{I}_d)$  denote the multivariate Gaussian distribution with mean  $m$  and covariance matrix  $\sigma^2 \mathbf{I}_d$ . If  $m_1, m_2 \in \mathbb{R}^d$  and  $\sigma > 0$ , then*

$$E_\gamma(\mathcal{N}(m_1, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(m_2, \sigma^2 \mathbf{I}_d)) = Q \left( \frac{\log \gamma}{\beta} - \frac{\beta}{2} \right) - \gamma Q \left( \frac{\log \gamma}{\beta} + \frac{\beta}{2} \right),$$

with  $\beta = \frac{\|m_2 - m_1\|}{\sigma}$  and  $Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} du$ .

This lemma motivates the following definition.

**Definition 1.** *For  $\gamma \geq 1$ , we let  $\theta_\gamma : [0, \infty) \rightarrow [0, 1]$  be the function defined by*

$$\begin{aligned} \theta_\gamma(r) &:= E_\gamma(\mathcal{N}(ru, \mathbf{I}_d) \| \mathcal{N}(0, \mathbf{I}_d)) \\ &= Q \left( \frac{\log \gamma}{r} - \frac{r}{2} \right) - \gamma Q \left( \frac{\log \gamma}{r} + \frac{r}{2} \right), \end{aligned}$$

where  $u \in \mathbb{R}^d$  is any vector of unit norm.

It is straightforward to verify that  $r \mapsto \theta_\gamma(r)$  is a non-decreasing mapping for every  $\gamma \geq 1$ . This intuitive property is useful to compute the contraction coefficient of the amplitude-constrained additive Gaussian kernel.

Given  $\mathcal{D} \subset \mathbb{R}^d$  and  $\sigma > 0$ , the  $\mathcal{D}$ -constrained additive Gaussian kernel  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathbb{R}^d)$  is defined by

$$K(\cdot | x) = \mathcal{N}(x, \sigma^2 \mathbf{I}_d).$$

Note that  $\mathcal{D} = \{x \in \mathbb{R}^d : \|x\|^2 \leq dA\}$  recovers the usual amplitude-constrained additive Gaussian kernel in (3). The following proposition is obtained by combining the previous lemma and our generalization of Dobrushin's formula in Theorem 2, as shown in Appendix B. Recall that

$$\text{dia}(\mathcal{D}) = \sup_{x_1, x_2 \in \mathcal{D}} \|x_2 - x_1\|.$$

**Proposition 1.** *Let  $\gamma \geq 1$ ,  $\mathcal{D} \subset \mathbb{R}^d$ , and  $\sigma > 0$ . If  $K$  is the  $\mathcal{D}$ -constrained additive Gaussian kernel, then*

$$\eta_\gamma(K) = \theta_\gamma \left( \frac{\text{dia}(\mathcal{D})}{\sigma} \right).$$

The contraction coefficient of the additive Gaussian kernel under  $E_\gamma$ -divergence is trivial<sup>1</sup>, i.e.,  $\eta_\gamma(K) = 1$ . Hence, Proposition 1 shows that constraints to the channel input must be imposed in order for the

<sup>1</sup>Recall that  $\eta_{\text{TV}}(K) = 1$  for any Gaussian channel  $K$  without input constraints [8] and, in addition,  $\eta_{\text{TV}}(K) = 1$  if and only if  $\eta_f(K) = 1$  for any non-linear function  $f$  [16, Prop. II.4.12].

contraction coefficient to be non-trivial. Given this proposition, Theorem 1 is a simple consequence of the data processing inequality for  $f$ -divergences. In Section IV we provide differential privacy guarantees for different noisy iterative algorithms based on these results.

**Remark 1.** Note that the  $E_\gamma$ -divergence between Gaussian distributions  $\mathcal{N}(m_1, \sigma^2 \mathbf{I}_d)$  and  $\mathcal{N}(m_2, \sigma^2 \mathbf{I}_d)$  can be concisely expressed as

$$E_\gamma(\mathcal{N}(m_1, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(m_2, \sigma^2 \mathbf{I}_d)) = \theta_\gamma \left( \frac{\|m_2 - m_1\|}{\sigma} \right). \quad (15)$$

The mapping  $r \mapsto \theta_\gamma(r)$  is closely related to the functions  $\theta(r)$  by Polyanskiy and Wu [8, Sec. 2.2],  $h(\eta)$  by Balle and Wang in [95, Lemma 7], and  $R_\alpha$  by Feldman et al. [73, Def. 10].

### B. A General Upper Bound for Output $f$ -divergence

The fundamental inequality in (13) implies that, for any probability measures  $\mu$  and  $\nu$ ,

$$D_f(\mu K \| \nu K) \leq \eta_{TV}(K) D_f(\mu \| \nu), \quad (16)$$

where  $f$  is a convex function with  $f(1) = 0$ . We end this section by proposing an improvement to this inequality which is very effective in combination with Theorem 2.

As established in [96, Proposition 3], if  $f$  is twice differentiable and convex, then  $D_f(\mu \| \nu)$  can be expressed in terms of the  $E_\gamma$ -divergence as

$$\begin{aligned} D_f(\mu \| \nu) &= \int_0^\infty f''(\gamma) E_\gamma(\mu \| \nu) d\gamma \\ &= \int_1^\infty [f''(\gamma) E_\gamma(\mu \| \nu) + \gamma^{-3} f''(\gamma^{-1}) E_\gamma(\nu \| \mu)] d\gamma. \end{aligned} \quad (17)$$

This expression and the definition of the contraction coefficient  $\eta_\gamma(K)$  directly yield the following corollary.

**Corollary 1.** Let  $f : (0, \infty) \rightarrow \mathbb{R}$  be a twice differentiable convex function with  $f(1) = 0$ . If  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  is a Markov kernel, then, for any probability measures  $\mu, \nu \in \mathcal{P}(\mathcal{D})$ ,

$$D_f(\mu K \| \nu K) \leq \int_1^\infty \eta_\gamma(K) [f''(\gamma) E_\gamma(\mu \| \nu) + \gamma^{-3} f''(\gamma^{-1}) E_\gamma(\nu \| \mu)] d\gamma.$$

Observe that, by (13),  $\eta_\gamma(K) \leq \eta_{TV}(K)$  for all  $\gamma > 0$ . Hence, the integral representation (17) shows that the bound in the previous corollary improves over (16). To illustrate the magnitude of this improvement, we consider the following two examples. Recall that  $\chi^2$ -divergence is the  $f$ -divergence with  $f(t) = (t - 1)^2$ .

*Example 1.* Let  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \mathcal{N}(2, 1)$ . In addition, let  $X_{\mathcal{D}}$  and  $Y_{\mathcal{D}}$  be the projections of  $X$  and  $Y$  onto the set

$$\mathcal{D} = \{x \in \mathbb{R} : |x| \leq 1/2\},$$

respectively, and denote by  $\mu$  and  $\nu$  their corresponding distributions.

Let  $K$  be the  $\mathcal{D}$ -constrained additive Gaussian kernel. By Proposition 1,  $\eta_{TV}(K) = \theta_1(1)$  and  $\eta_\gamma(K) = \theta_\gamma(1)$ . Thus, (16) implies that

$$\begin{aligned} \chi^2(\mu K \| \nu K) &\leq \eta_{TV}(K) \chi^2(\mu \| \nu) \\ &= 2\theta_1(1) \int_1^\infty E_\gamma(\mu \| \nu) + \gamma^{-3} E_\gamma(\nu \| \mu) d\gamma, \end{aligned}$$

where the equality follows from (17). By the data processing inequality,

$$\begin{aligned} \mathsf{E}_\gamma(\mu\|\nu) &\leq \mathsf{E}_\gamma(\mathcal{N}(0,1)\|\mathcal{N}(2,1)) = \theta_\gamma(2), \\ \mathsf{E}_\gamma(\nu\|\mu) &\leq \mathsf{E}_\gamma(\mathcal{N}(2,1)\|\mathcal{N}(0,1)) = \theta_\gamma(2). \end{aligned}$$

Therefore, Lemma 3 implies that

$$\begin{aligned} \chi^2(\mu K\|\nu K) &\leq 2\theta_1(1) \int_1^\infty (1+\gamma^{-3})\theta_\gamma(2) d\gamma \\ &= 0.49. \end{aligned}$$

Note, however, that Corollary 1 implies that

$$\chi^2(\mu K\|\nu K) \leq 2 \int_1^\infty \theta_\gamma(1) [\mathsf{E}_\gamma(\mu\|\nu) + \gamma^{-3}\mathsf{E}_\gamma(\nu\|\mu)] d\gamma.$$

As before, the data processing inequality and Lemma 3 lead to

$$\begin{aligned} \chi^2(\mu K\|\nu K) &\leq 2 \int_1^\infty (1+\gamma^{-3})\theta_\gamma(1)\theta_\gamma(2) d\gamma \\ &= 0.26. \end{aligned}$$

In conclusion, the bound for  $\chi^2(\mu K\|\nu K)$  obtained from (16) is almost twice the corresponding bound obtained from Corollary 1.

*Example 2.* Let  $K$  be the binary-input binary-output channel with crossover probabilities  $a, b \in [0, \frac{1}{2}]$ , i.e.,

$$K = \begin{bmatrix} \bar{a} & a \\ b & \bar{b} \end{bmatrix}.$$

Theorem 2 implies that, for all  $\gamma \geq 1$ ,

$$\eta_\gamma(K) = \max\{(\bar{a} - \gamma b)^+, (\bar{b} - \gamma a)^+\},$$

and, in particular,

$$\eta_{TV}(K) = 1 - a - b.$$

Thus, (16) implies that, for any binary probability measures  $\mu$  and  $\nu$ ,

$$\begin{aligned} \chi^2(\mu K\|\nu K) &\leq (1-a-b)\chi^2(\mu\|\nu) \\ &= 2(1-a-b) \int_1^\infty [\mathsf{E}_\gamma(\mu\|\nu) + \gamma^{-3}\mathsf{E}_\gamma(\nu\|\mu)] d\gamma, \end{aligned}$$

where the equality comes from the integral representation (17). In particular, for  $a = 0.1$ ,  $b = 0.4$ ,  $\mu = \text{Bernoulli}(0.1)$ , and  $\nu = \text{Bernoulli}(0.4)$ , we obtain

$$\chi^2(\mu K\|\nu K) \leq 0.19.$$

On the other hand, Corollary 1 renders

$$\chi^2(\mu K\|\nu K) \leq 2 \int_1^{\frac{\bar{b}}{a} \vee \frac{\bar{a}}{b}} [(\bar{a} - \gamma b)^+ \vee (\bar{b} - \gamma a)^+] [\mathsf{E}_\gamma(\mu\|\nu) + \gamma^{-3}\mathsf{E}_\gamma(\nu\|\mu)] d\gamma.$$

For our particular choice of  $a$ ,  $b$ ,  $\mu$ , and  $\nu$ , we obtain

$$\chi^2(\mu K\|\nu K) \leq 0.17.$$

Hence, as in the previous example, Corollary 1 produces a tighter estimate for  $\chi^2(\mu K\|\nu K)$  than (16).

### III. LDP AS THE CONTRACTION OF $E_\gamma$ -DIVERGENCE

In this section we establish the equivalence between the notions of local differential privacy (LDP) and contraction under  $E_\gamma$ -divergence. We apply this equivalence in a plug-and-play manner to obtain estimates for the cost of privacy in the minimax risk, Bayesian risk, and binary hypothesis testing settings. Motivated by the equivalence between LDP and contraction under  $E_\gamma$ -divergence, we obtain these estimates by bounding the cost of privacy directly in terms of  $E_\gamma$ -divergence.

A privacy mechanism is a (potentially random) mapping from a given set  $\mathcal{D}$  into another set  $\mathcal{W}$ . As a result, it is customary to represent privacy mechanisms as Markov kernels  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$ . In this context, LDP can be expressed as follows.

**Definition 2** ([40, 47]). *Let  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ . A privacy mechanism  $K : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{W})$  is  $(\varepsilon, \delta)$ -LDP if*

$$\sup_{x_1, x_2 \in \mathcal{D}} \sup_{A \subset \mathcal{W}} [K(A|x_1) - e^\varepsilon K(A|x_2)] \leq \delta.$$

We let  $\mathcal{Q}_{\varepsilon, \delta}$  be the family of all Markov kernels that are  $(\varepsilon, \delta)$ -LDP. An  $(\varepsilon, 0)$ -LDP mechanism is called  $\varepsilon$ -LDP.

Observe that, by (7), a Markov kernel  $K$  is  $(\varepsilon, \delta)$ -LDP if and only if

$$\sup_{x_1, x_2 \in \mathcal{D}} E_\gamma(K(\cdot|x_1) \| K(\cdot|x_2)) \leq \delta.$$

Thus, an immediate application of Theorem 2 shows that  $(\varepsilon, \delta)$ -LDP is equivalent to contraction under  $E_\gamma$ -divergence.

**Theorem 3.** *A mechanism  $K$  is  $(\varepsilon, \delta)$ -LDP if and only if  $\eta_{e^\varepsilon}(K) \leq \delta$ , i.e.,*

$$E_{e^\varepsilon}(\mu K \| \nu K) \leq \delta E_{e^\varepsilon}(\mu \| \nu), \quad \forall \mu, \nu.$$

We note that Duchi et al. [20] showed that if  $K$  is  $\varepsilon$ -LDP then

$$D_{KL}(\mu K \| \nu K) \leq 2(e^\varepsilon - 1)^2 TV^2(\mu, \nu). \quad (18)$$

From this result, they concluded that  $\varepsilon$ -LDP acts as a contraction on the space of probability measures. Theorem 3 makes this observation precise by showing that  $\varepsilon$ -LDP is in fact equivalent to contraction under  $E_{e^\varepsilon}$ -divergence. Indeed, according to Theorem 3, a privacy mechanism  $K$  is  $\varepsilon$ -LDP if and only if  $E_{e^\varepsilon}(\mu K \| \nu K) = 0$  for any distributions  $\mu$  and  $\nu$ . An example of a Markov kernel satisfying the latter property is the randomized response mechanism, as described next.

*Example 3.* Let  $\mathcal{D} = \mathcal{W} = \{0, 1\}$  and  $\varepsilon \geq 0$ . The randomized response mechanism [97], denoted by  $K_{RR}^\varepsilon$ , is the mechanism implemented by the binary symmetric channel with crossover probability  $\omega_\varepsilon := (1 + e^\varepsilon)^{-1}$ . It is known that  $K_{RR}^\varepsilon$  is  $\varepsilon$ -LDP. For the sake of illustration, we verify this fact below using Theorem 3.

Let  $\mu = \text{Bernoulli}(p)$  and  $\nu = \text{Bernoulli}(q)$  with  $p, q \in [0, 1]$ . Observe that  $\mu K_{RR}^\varepsilon = \text{Bernoulli}(p * \omega_\varepsilon)$  and  $QK_{RR}^\varepsilon = \text{Bernoulli}(q * \omega_\varepsilon)$  where

$$a * b := a(1 - b) + (1 - a)b.$$

It is straightforward to verify that, for any  $p, q$ ,

$$\begin{aligned} p * \omega_\varepsilon - e^\varepsilon q * \omega_\varepsilon &\leq 0, \\ 1 - p * \omega_\varepsilon - e^\varepsilon (1 - q * \omega_\varepsilon) &\leq 0. \end{aligned}$$

The definition of the  $E_\gamma$ -divergence (6) implies that

$$E_{e^\varepsilon}(\mu K_{RR}^\varepsilon \| \nu K_{RR}^\varepsilon) = 0.$$

Thus, by Theorem 3, we conclude that  $K_{\text{RR}}^\varepsilon$  is  $\varepsilon$ -LDP.

A generalization of this mechanism for  $|\mathcal{D}| = k \geq 2$  has been reported in the literature (see, e.g., [42, 98]). Specifically, the so-called  $k$ -ary randomized response mechanism  $K_{k\text{RR}}^\varepsilon : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{D})$  is defined by

$$K_{k\text{RR}}^\varepsilon(z|x) = \begin{cases} \frac{e^\varepsilon}{k-1+e^\varepsilon} & z = x, \\ \frac{1}{k-1+e^\varepsilon} & z \neq x. \end{cases}$$

Using an argument similar to the one in the paragraph above, the reader can verify that  $K_{k\text{RR}}^\varepsilon$  is  $\varepsilon$ -LDP.

For each  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ , we define

$$\varphi(\varepsilon, \delta) := 1 - (1 - \delta)e^{-\varepsilon}.$$

By combining (14) and Theorem 3, we conclude that if  $K$  is  $(\varepsilon, \delta)$ -LDP, i.e.,  $K \in \mathcal{Q}_{\varepsilon, \delta}$ , then

$$D_f(\mu K \| \nu K) \leq \varphi(\varepsilon, \delta) D_f(\mu \| \nu) \quad \forall \mu, \nu. \quad (19)$$

This direct consequence of Theorem 3 reveals a remarkable structural property of LDP mechanisms: any LDP mechanism is contractive under *all*  $f$ -divergences.

In the sequel we focus on two multi-user settings which are common in the literature. Assume there are  $n$  users, each in possession of a datapoint  $X_i \in \mathcal{X}$ ,  $i \in [n]$ . Furthermore, assume that the users wish to apply a mechanism  $K_i$  that generates a privatized version of  $X_i$ , denoted by  $Z_i \in \mathcal{Z}_i$ .

1) *Non-interactive setting*: The collection of mechanisms  $\{K_i\}$  is said to be *non-interactive* if the distribution of  $Z_i$  is entirely determined by  $X_i$  and independent of  $(X_j, Z_j)$  for  $j \neq i$ . When all users apply the same mechanism  $K$ , we can view  $Z^n := (Z_1, \dots, Z_n)$  as independent applications of  $K$  to each  $X_i$ . In this case, the overall mechanism is the  $n$ -fold tensor power of  $K$ , denoted by  $K^{\otimes n}$ .

2) *Sequentially interactive setting*: The collection of mechanisms  $\{K_i\}$  is said to be *sequentially interactive* [20] if the distribution of  $Z_i$  depends on  $X_i$ , the datapoint in possession of user  $i$ , and  $Z_1, \dots, Z_{i-1}$ , the output of the  $i-1$  previous mechanisms. Note that  $K_i$  is a Markov kernel with domain  $\mathcal{D} = \mathcal{X} \times \mathcal{Z}_1 \times \dots \times \mathcal{Z}_{i-1}$ . In this case, we denote the overall mechanism by  $K^n$ .

Next, we extend (19) for non-interactive mechanisms. For each  $n \in \mathbb{N}$ ,  $\varepsilon \geq 0$ , and  $\delta \in [0, 1]$ , we define

$$\varphi_n(\varepsilon, \delta) := 1 - e^{-n\varepsilon}(1 - \delta)^n.$$

Fix an  $(\varepsilon, \delta)$ -LDP mechanism  $K$  and consider the corresponding non-interactive mechanism  $K^{\otimes n}$ . By combining Lemma 2 and Theorem 3, we obtain that

$$\eta_f(K^{\otimes n}) \leq \varphi_n(\varepsilon, \delta),$$

where  $f$  is any convex function with  $f(1) = 0$ . From this inequality and (14), we conclude that

$$D_f(\mu K^{\otimes n} \| \nu K^{\otimes n}) \leq \varphi_n(\varepsilon, \delta) D_f(\mu \| \nu) \quad \forall \mu, \nu, \quad (20)$$

which generalizes (19) as desired. In the rest of this section we rely on the results derived so far to estimate the cost of local differential privacy in some statistical settings.

#### A. Private Minimax Risk

Let  $\mathcal{P} \subseteq \mathcal{P}(\mathcal{X})$  be a family of distributions, let  $\vartheta$  be a parameter space, and let  $\theta : \mathcal{P} \rightarrow \vartheta$  be a function assigning a parameter to each distribution in  $\mathcal{P}$ . Also, let  $X^n = (X_1, \dots, X_n)$  be independent and identically distributed (i.i.d.) samples drawn from a distribution  $P$  with parameter  $\theta(P)$ . We assume that each user possesses a sample  $X_i$  and applies a sequentially interactive privacy-preserving mechanism  $K_i$  to obtain  $Z_i$ . Given the sequences  $\{Z_i\}_{i=1}^n$ , the goal is to estimate  $\theta(P)$  through an estimator  $\Psi : \mathcal{Z}^n \rightarrow \vartheta$ .

The quality of such estimator is assessed by a semi-metric  $\ell : \vartheta \times \vartheta \rightarrow \mathbb{R}_+$  and is used to define the private minimax risk

$$\mathcal{R}_n(\mathcal{P}, \ell, \varepsilon, \delta) := \inf_{\mathsf{K}_i \in \mathcal{Q}_{\varepsilon, \delta}} \inf_{\Psi} \sup_{P \in \mathcal{P}} \mathbb{E}[\ell(\Psi(Z^n), \theta(P))],$$

where the first infimum is taken over all  $\mathsf{K}_1, \dots, \mathsf{K}_n$  which are  $(\varepsilon, \delta)$ -LDP. The quantity  $R_n(\mathcal{P}, \ell, \varepsilon, \delta)$  uniformly characterizes the optimal rate of private statistical estimation over the family  $\mathcal{P}$  using the best possible estimator and privacy-preserving mechanisms in  $\mathcal{Q}_{\varepsilon, \delta}$ . In the absence of privacy constraints (i.e.,  $Z^n = X^n$ ), we denote the minimax risk by  $\mathcal{R}_n(\mathcal{P}, \ell)$ .

A typical first step in deriving information-theoretic lower bounds for the minimax risk is to reduce the above estimation problem to a testing problem via Le Cam's, Fano's, or Assouad's method [99–101]. For ease of exposition, in the sequel we focus only on Le Cam's method, which relies on binary hypothesis testing, although a similar reasoning could be applied to multiple hypothesis testing settings (e.g., Fano's and Assouad's methods). The canonical binary hypothesis testing problem is defined as follows: Nature chooses a random variable  $V$  uniformly at random from  $\{0, 1\}$ , then, conditioned on  $V = v$ , the samples  $X^n$  are drawn i.i.d. from  $P_v \in \mathcal{P}$ , denoted by  $X^n \sim P_v^{\otimes n}$ . It is well-known [99–101] that if  $\ell(\theta(P_0), \theta(P_1)) \geq 2\tau$  for some  $\tau > 0$ , then

$$\mathcal{R}_n(\mathcal{P}, \ell) \geq \tau \mathsf{P}_e(V|X^n),$$

where  $\mathsf{P}_e(V|X^n)$  denotes the probability of error in guessing  $V$  given  $X^n$ . In its simplest form, Le Cam's method relies on the inequality

$$\mathsf{P}_e(V|X^n) \geq \frac{1}{2} [1 - \mathsf{TV}(P_0^{\otimes n}, P_1^{\otimes n})],$$

see, e.g., [99, Lemma 1] or [101, Theorem 2.2], which yields the following lower bound for the minimax risk

$$\mathcal{R}_n(\mathcal{P}, \ell) \geq \frac{\tau}{2} [1 - \mathsf{TV}(P_0^{\otimes n}, P_1^{\otimes n})] \tag{21}$$

$$\geq \frac{\tau}{2} \left[ 1 - \sqrt{\frac{n}{2} D_{\text{KL}}(P_0 \| P_1)} \right], \tag{22}$$

where the second inequality follows from Pinsker's inequality<sup>2</sup> and the chain rule for KL-divergence.

In the presence of privacy, the estimator  $\Psi$  depends on  $Z^n$  which is generated upon  $X^n$  by a sequentially interactive mechanism  $\mathsf{K}^n$ . To derive the private counterpart of (21), we need to replace  $P_v^{\otimes n}$ , the marginal distribution of  $X^n$  conditioned on  $V = v$ , with  $P_v^{\otimes n} \mathsf{K}^n$ , the marginal distribution of  $Z^n$  conditioned on  $V = v$ . A lower bound for  $\mathcal{R}_n(\mathcal{P}, \ell, \varepsilon, \delta)$  is therefore obtained by deriving an upper bound for  $\mathsf{TV}(P_0^{\otimes n} \mathsf{K}^n, P_1^{\otimes n} \mathsf{K}^n)$  for all  $\mathsf{K}^n \subset \mathcal{Q}_{\varepsilon, \delta}$ . This strategy is implemented in the following lemma, whose proof could be found in Appendix C.

**Lemma 4.** *If  $P_0, P_1 \in \mathcal{P}$  satisfy  $\ell(\theta(P_0), \theta(P_1)) \geq 2\tau$  for some  $\tau > 0$ , then*

$$\mathcal{R}_n(\mathcal{P}, \ell, \varepsilon, \delta) \geq \frac{\tau}{2} \left[ 1 - \sqrt{\frac{\varphi(\varepsilon, \delta)n}{2} D_{\text{KL}}(P_0 \| P_1)} \right],$$

<sup>2</sup>Observe that Pinsker's inequality is ineffective when  $D_{\text{KL}}(P_0 \| P_1)$  is sufficiently large. In that case, Pinsker's inequality could be replaced by the Bretagnolle-Huber inequality [102]

$$\mathsf{TV}(P_0, P_1) \leq \sqrt{1 - e^{-D_{\text{KL}}(P_0 \| P_1)}},$$

or Vajda's inequality [103]

$$\log \left( \frac{1 + \mathsf{TV}(P_0, P_1)}{1 - \mathsf{TV}(P_0, P_1)} \right) - \frac{2\mathsf{TV}(P_0, P_1)}{1 + \mathsf{TV}(P_0, P_1)} \leq D_{\text{KL}}(P_0 \| P_1).$$

where  $\varphi(\varepsilon, \delta) := 1 - (1 - \delta)e^{-\varepsilon}$ .

By comparing the previous lemma with the original non-private Le Cam's method (22), we observe that the effect of  $(\varepsilon, \delta)$ -LDP is to reduce the effective sample size from  $n$  to  $(1 - (1 - \delta)e^{-\varepsilon})n$ . Setting  $\delta = 0$  and  $\varepsilon < 0.224$ , this result strengthens Duchi et al. [20, Corollary 2], where the effective sample size was shown to be  $4\varepsilon^2 n$  for sufficiently small  $\varepsilon$ .

*Example 4. (1-dimensional mean estimation)* For some  $k > 1$ , we assume that  $\mathcal{P} = \mathcal{P}_k$  is given by

$$\mathcal{P}_k := \left\{ P \in \mathcal{P}(\mathcal{X}) : |\mathbb{E}_P[X]| \leq 1, \mathbb{E}_P[|X|^k] \leq 1 \right\}.$$

Consider the problem of estimating  $\theta(P) = \mathbb{E}_P[X]$  when  $\ell = \ell_2^2$ , i.e., the loss is given by the squared  $\ell_2$  metric. This problem was first studied by Duchi et al. in [20, Prop. 1] where it was shown that  $\mathcal{R}_n(\mathcal{P}_k, \ell_2^2, \varepsilon, 0) \geq (\varepsilon^2 n)^{-(k-1)/k}$  for  $\varepsilon \leq 1$ . Applying our framework to this example, we obtain a similar lower bound that holds for all  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ , stated below.

**Corollary 2.** *For the 1-dimensional mean estimation problem, for all  $k > 1$ ,  $\varepsilon \geq 0$ , and  $\delta \in (0, 1)$ ,*

$$\mathcal{R}_n(\mathcal{P}_k, \ell_2^2, \varepsilon, \delta) \gtrsim \min \left\{ 1, [\varphi^2(\varepsilon, \delta)n]^{-\frac{k-1}{k}} \right\}. \quad (23)$$

It is worth instantiating this corollary, whose proof can be found in Appendix D, for some special values of  $k$ . Consider first the usual finite variance setting, i.e.,  $k = 2$ . In the non-private case, it is known that the sample mean has mean-squared error that scales as  $n^{-1}$ . According to Corollary 2, this rate worsens to  $(\varphi(\varepsilon, \delta)\sqrt{n})^{-1}$  in the presence of an  $(\varepsilon, \delta)$ -LDP requirement. Now, consider the limiting setting in which  $k \rightarrow \infty$ . Observe that the moment condition  $\mathbb{E}_p[|X|^k] \leq 1$  implies the boundedness of  $X$ . In this case, the non-private minimax risk scales as  $n^{-1}$ , while Corollary 2 implies that its LDP counterpart scales as  $(\varphi^2(\varepsilon, \delta)n)^{-1}$ .

We end this section with an alternative to Lemma 4 in the non-interactive setting. Its proof, which can be found in Appendix E, relies on Lemma 2 and leads to tighter bounds for certain values of  $n$ ,  $\varepsilon$ , and  $\delta$ .

**Lemma 5.** *If  $P_0, P_1 \in \mathcal{P}$  satisfy  $\ell(\theta(P_0), \theta(P_1)) \geq 2\tau$  for some  $\tau > 0$ , then*

$$\mathcal{R}_n(\mathcal{P}, \ell, \varepsilon, \delta) \geq \frac{\tau}{2} \left[ 1 - \sqrt{\frac{\varphi_n^2(\varepsilon, \delta)n}{2} D_{\text{KL}}(P_0 \| P_1)} \right],$$

where  $\varphi_n(\varepsilon, \delta) := 1 - e^{-n\varepsilon}(1 - \delta)^n$ .

## B. Private Bayesian Risk

In the minimax setting, the worst-case parameter is considered which usually leads to over-pessimistic bounds. In practice, the parameter that incurs a worst-case risk may appear with very small probability. To capture this prior knowledge, it is reasonable to assume that the true parameter is sampled from an underlying prior distribution. In this case, we are interested in the *Bayesian risk* of the problem as described next. Let  $\vartheta$  be a parameter space endowed with a prior distribution  $P_\Theta$  and let  $\mathcal{P} = \{P_{X|\Theta}(\cdot|\theta) : \theta \in \vartheta\}$  be a collection of parametric probability distributions over  $\mathcal{X}$ . Given an i.i.d. sequence  $X^n$  drawn from  $P_{X|\Theta}$ , the goal is to estimate  $\Theta$  from a privatized sequence  $Z^n$  via an estimator  $\Psi : \mathcal{Z}^n \rightarrow \vartheta$ . Throughout this section, we focus on the non-interactive setting. In this case, the private Bayesian risk is defined as

$$R_n^{\text{Bayes}}(P_\Theta, \ell, \varepsilon, \delta) := \inf_{K \in \mathcal{Q}_{\varepsilon, \delta}} \inf_{\Psi} \mathbb{E}[\ell(\Theta, \Psi(Z^n))], \quad (24)$$

where the expectation is taken with respect to the randomness of both  $\Theta \sim P_\Theta$  and  $Z^n$ . It is evident that  $R_n^{\text{Bayes}}(P_\Theta, \ell, \varepsilon, \delta)$  must depend on the prior  $P_\Theta$ . One way to quantify this dependence is through the so-called *small ball probability* of  $\Theta$  with respect to  $\ell$  defined as

$$\mathcal{L}(\zeta) := \sup_{\theta \in \vartheta} \Pr(\ell(\Theta, \theta) \leq \zeta).$$

Xu and Raginsky [18] showed that the non-private Bayesian risk ( $Z^n = X^n$ ), denoted by  $R_n^{\text{Bayes}}(P_\Theta, \ell)$ , is lower bounded as

$$R_n^{\text{Bayes}}(P_\Theta, \ell) \geq \sup_{\zeta > 0} \zeta \left[ 1 - \frac{I(\Theta; X^n) + \log 2}{\log(1/\mathcal{L}(\zeta))} \right]. \quad (25)$$

By replacing  $I(\Theta; X^n)$  with  $I(\Theta; Z^n)$  in the previous inequality, an application of (20) directly leads to the following lower bound for  $R_n^{\text{Bayes}}(P_\Theta, \ell, \varepsilon, \delta)$ .

**Corollary 3.** *In the non-interactive setting, the private Bayesian risk  $R_n^{\text{Bayes}}(P_\Theta, \ell, \varepsilon, \delta)$  is bounded below by*

$$\sup_{\zeta > 0} \zeta \left[ 1 - \frac{\varphi_n(\varepsilon, \delta) I(\Theta; X^n) + \log 2}{\log(1/\mathcal{L}(\zeta))} \right], \quad (26)$$

where  $\varphi_n(\varepsilon, \delta) := 1 - e^{-n\varepsilon}(1 - \delta)^n$ .

The following theorem, whose proof can be found in Appendix F, provides a lower bound for the private Bayesian risk that directly involves  $E_\gamma$ -divergence, leading to a tighter bound than (26). For  $\gamma \geq 0$  and any pair of random variables  $(A, B) \sim P_{AB}$  with marginals  $P_A$  and  $P_B$ , we define their  $E_\gamma$ -information as

$$I_\gamma(A; B) := E_\gamma(P_{AB} \| P_A P_B).$$

**Theorem 4.** *In the non-interactive setting, the private Bayesian risk  $R_n^{\text{Bayes}}(P_\Theta, \ell, \varepsilon, \delta)$  is bounded below by*

$$\sup_{\zeta > 0} \zeta [1 - \varphi_n(\varepsilon, \delta) I_{e^\varepsilon}(\Theta; X^n) - e^\varepsilon \mathcal{L}(\zeta)].$$

Furthermore, for  $n = 1$  we have that

$$R_1^{\text{Bayes}}(P_\Theta, \ell, \varepsilon, \delta) \geq \sup_{\zeta > 0} \zeta [1 - \delta I_{e^\varepsilon}(\Theta; X) - e^\varepsilon \mathcal{L}(\zeta)].$$

We compare the bound in Theorem 4 with those of Corollary 3 in the next example.

*Example 5.* Suppose  $\Theta$  is uniformly distributed on  $[0, 1]$ ,  $P_{X|\Theta=\theta} = \text{Bernoulli}(\theta)$ , and  $\ell(\theta, \theta') = |\theta - \theta'|$ . Observe that in this case  $\mathcal{L}(\zeta) \leq \min\{2\zeta, 1\}$ . For ease of notation, let  $\gamma = e^\varepsilon$ . Note that

$$I_\gamma(\Theta; X^n) = \int_0^1 E_\gamma(P_{X^n|\Theta=\theta} \| P_{X^n}) d\theta.$$

A routine calculation shows that, for any  $\theta \in [0, 1]$ ,

$$\begin{aligned} P_{X^n|\Theta=\theta}(x^n) &= \theta^{s(x^n)} (1 - \theta)^{n-s(x^n)}, \\ P_{X^n}(x^n) &= \frac{s(x^n)! (n - s(x^n))!}{(n + 1)!}, \end{aligned}$$

where  $s(x^n)$  is the number of 1's in  $x^n$ . Given these marginal and conditional distribution, one can obtain that

$$I_\gamma(\Theta; X^n) = \sum_{s=0}^n \int_0^1 \left[ \frac{n! \theta^s (1 - \theta)^{n-s}}{s!(n-s)!} - \frac{\gamma}{n+1} \right]_+ d\theta.$$

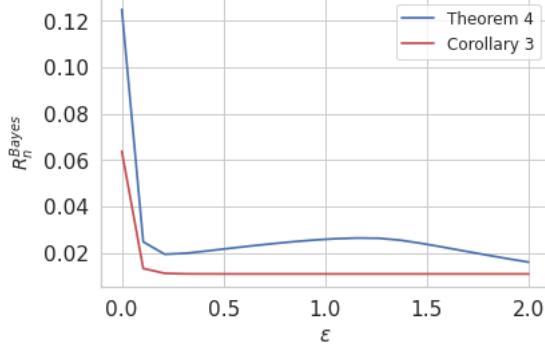


Fig. 1. Comparison of the lower bounds obtained from Theorem 4 and the private version of [18, Theorem 1] described in Corollary 3 for Example 5 with  $\delta = 10^{-4}$  and  $n = 20$ .

Plugging this equation into Theorem 4, we arrive at a maximization problem that can be numerically solved. Similarly, we compute

$$I(\Theta; X^n) = \int_0^1 D_{\text{KL}}(P_{X^n|\Theta=\theta} \| P_{X^n}) d\theta,$$

plug it into Corollary 26, and numerically solve the resulting optimization problem. In Figure 1, we compare these two lower bounds for  $\delta = 10^{-4}$  and  $n = 20$ . Observe the significant advantage of Theorem 4 against Corollary 26 for small  $\epsilon$ .

**Remark 2.** Although we are mainly interested in the private Bayesian risk, it is possible to obtain from the proof of Theorem 4 that the non-private Bayesian risk  $R_n^{\text{Bayes}}(P_\Theta, \ell)$  is bounded below by

$$\sup_{\gamma \geq 0} \sup_{\zeta > 0} \zeta [1 - I_\gamma(\Theta; X^n) - \gamma \mathcal{L}(\zeta) - (1 - \gamma)_+]. \quad (27)$$

In order to compare the previous bound with its non-private counterpart (25), we consider the following example. Suppose  $\Theta$  is uniformly distributed on  $[0, 1]$ ,  $P_{X|\Theta=\theta} = \text{Bernoulli}(\theta)$ , and  $\ell(\theta, \theta') = |\theta - \theta'|$ . It can be shown that  $I(\Theta; X) = 0.19$  nats while

$$I_\gamma(\Theta; X) = \begin{cases} 0.25\gamma^2 & \text{if } \gamma \in [0, 1], \\ 0.25(\gamma - 2)^2 & \text{if } \gamma \in [1, 2], \\ 0 & \text{otherwise.} \end{cases}$$

As in Example 5, it can be verified that (25) produces the lower bound  $R_1^{\text{Bayes}}(P_\Theta, \ell_1) \geq 0.03$ , whereas our bound (27) yields  $R_1^{\text{Bayes}}(P_\Theta, \ell_1) \geq 0.08$ .

### C. Private Binary Hypothesis Testing

We now turn our attention to the well-known problem of binary hypothesis testing under local differential privacy constraints. Assume that we observe  $n$  i.i.d. samples  $X^n$  drawn from a distribution  $Q \in \mathcal{P}(\mathcal{X})$ . We assume further that each  $X_i$  is mapped to  $Z_i$  via a privacy-preserving mechanism  $K \in \mathcal{Q}_{\epsilon, \delta}$ , i.e., we assume a non-interactive setting with  $K_i = K$ . Given  $Z^n$ , the goal is to distinguish between the null hypothesis  $H_0 : Q = P_0$  and the alternative hypothesis  $H_1 : Q = P_1$ . Let  $T$  be a binary statistic generated from a randomized decision rule  $P_{T|Z^n} : \mathcal{Z}^n \rightarrow \mathcal{P}(\{0, 1\})$ , where  $T = 1$  indicates that  $H_0$  is rejected. The type I and type II error probabilities corresponding to this statistic are given by  $\Pr(T = 1 | H_0)$  and

$\Pr(T = 0|H_1)$ , respectively. To capture the optimal tradeoff between type I and type II error probabilities, it is customary to define

$$\beta_n^{\varepsilon, \delta}(\alpha) := \inf_{K \in \mathcal{Q}_{\varepsilon, \delta}} \inf_{\substack{P_{T|Z^n}: \\ \Pr(T=1|H_0) \leq \alpha}} \Pr(T = 0|H_1),$$

The following corollary, whose proof relies on (19) and can be found in Appendix G, provides an asymptotic lower bound for  $\beta_n^{\varepsilon, \delta}(\alpha)$ .

**Corollary 4.** *For any  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ , we have that*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \beta_n^{\varepsilon, \delta}(\alpha) \geq -\varphi(\varepsilon, \delta) D_{\text{KL}}(P_0 \| P_1). \quad (28)$$

Kairouz et al. [42, Sec. 3] proved a similar result which is optimal for sufficiently small (albeit unspecified)  $\varepsilon$  and  $\delta = 0$ . Recall that the Chernoff-Stein lemma [104, Thm. 11.8.3] establishes that  $D_{\text{KL}}(P_0 \| P_1)$  is the asymptotic exponential decay rate of  $\beta_n(\alpha)$  in the absence of privacy constraints. Thus, the above corollary exhibits, once again, a reduction of the effective sample size from  $n$  to  $\varphi(\varepsilon, \delta)n$  in the presence of an  $(\varepsilon, \delta)$ -LDP constraint.

We end this section with a remark. In the spirit of Theorem 4, it is possible to investigate  $\alpha \mapsto \beta_n^{\varepsilon, \delta}(\alpha)$  directly in terms of  $E_\gamma$ -divergence, for instance, by considering the dual representation of  $E_\gamma$ -divergence. This method can potentially lead to a tighter lower bound than Corollary 4 and is left as future work. The details of this approach for the non-private case can be found in the pioneering work of [13].

#### IV. DIFFERENTIALLY PRIVATE ONLINE LEARNING

In this section, we use Theorem 2 to develop a framework for quantifying the DP guarantees of general online learning algorithms. Before delving into the technical results, we first describe the framework of online learning studied in this section and give the definition of central differential privacy.

##### A. Online Learning Algorithms

Let  $\mathcal{W} \subset \mathbb{R}^d$  denote a parameter space, e.g., the coefficients of a linear regression model, and let  $\Pi_{\mathcal{W}}(\cdot)$  denote the projection operator onto  $\mathcal{W}$ . We describe typical *online learning* algorithms in the next definition.

**Definition 3** (Online Algorithm). *An online learning algorithm  $\mathcal{M}$  proceeds as follows. A learner initiates the algorithm by taking a random point  $W_1$  from  $\mathcal{W}$ . Once  $W_1$  is chosen, a convex cost function  $f_1 : \mathcal{W} \rightarrow \mathbb{R}$  is revealed, implying that the cost associated with  $W_1$  is  $f_1(W_1)$ . Upon observing  $f_t$  and  $W_t$ , the learner at time  $t + 1$  chooses  $W_{t+1}$  according to an update rule  $\Psi_{f_t} : \mathcal{W} \rightarrow \mathcal{W}$ , specifically,*

$$W_{t+1} = \Pi_{\mathcal{W}}(\Psi_{f_t}(W_t)).$$

After  $n$  iterations, the algorithm outputs  $W_{n+1}$ . We say that  $\mathcal{M}$  is randomized if, for  $\{f_t\}$  and  $W_t$  fixed,  $W_{t+1}$  is a random variable on  $\mathcal{W}$ .

Letting  $\mathcal{F}$  be the collection of all possible convex cost functions, an online learning algorithm can thus be viewed as the mapping

$$\mathcal{M} : \mathcal{F}^n \rightarrow \mathcal{W},$$

given by

$$\mathcal{M}(\{f_1, \dots, f_n\}) = W_{n+1}.$$

For brevity, we denote  $\{f_1, \dots, f_n\}$  by  $\{f_t\}$ . It is worth noting that in this setting we assume that the collection of cost functions  $\{f_1, \dots, f_n\}$  is fixed before the algorithm is started—often referred to as the

*oblivious setting* in the literature [49, Section 5.5]. We also assume that  $\Psi_{f_t}$ , the update function at time  $t$ , only depends on  $f_t$  and not on the previous cost functions  $f_1, \dots, f_{t-1}$ .

The goal of the learner is to minimize the *regret*, i.e., the difference between the cumulative cost of the algorithm's choices and that of the best fixed (offline) solution in hindsight. Specifically, the regret of a randomized algorithm  $\mathcal{M}$  after  $n$  iterations is defined as

$$\mathbb{E} \left[ \sum_{t=1}^n f_t(W_t) \right] - \min_{w \in \mathcal{W}} \sum_{t=1}^n f_t(w), \quad (29)$$

where the expectation is taken over the algorithm's randomness. Typically, for online learning algorithms, a *sublinear* regret is sought, as it implies that asymptotically an online algorithm performs almost as well as the optimal solution in hindsight.

In many practical scenarios, the cost functions  $\{f_t\}$  might leak private information about the learner. For example, in offline setting (cf. Section IV-B)  $f_t(w) = \ell(w, x_t)$  where  $x_t$  is a ground truth observation (e.g., stock price) at time  $t$  and  $\ell$  is loss function. If the loss function is linear, then  $f_t(w)$  reveals significant information about the observation  $x_t$ . Other examples of privacy leakage in online algorithms are listed in [105, 106]. As such, we focus on privacy attacks aimed to infer information about the cost function  $f_i$  for some  $i \in [n]$  upon observing the output of the algorithm, i.e.,  $\mathcal{M}(\{f_t\}) = W_{n+1}$ . Thus, our goal is to design online algorithms such that  $W_{n+1}$  does not reveal significant information about any single cost function in  $\{f_t\}$ . The following definition formalizes this goal. We say that two collection of cost functions  $\{f_t\}$  and  $\{f'_t\}$  are *neighboring* at index  $i \in [n]$ , if  $f_t = f'_t$  for all  $t \in [n] \setminus i$  and  $f_i \neq f'_i$ . We denote this by  $\{f_t\} \stackrel{i}{\sim} \{f'_t\}$ .

**Definition 4** ([73]). *Given  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ , a randomized online algorithm  $\mathcal{M}$  is said to be  $(\varepsilon, \delta)$ -DP at index  $i \in [n]$ , if*

$$\sup_{\{f_t\} \stackrel{i}{\sim} \{f'_t\}} \mathsf{E}_{e^\varepsilon}(\mu_{n+1} \| \mu'_{n+1}) \leq \delta,$$

where  $\mu_{n+1}$  and  $\mu'_{n+1}$  are the distributions of  $\mathcal{M}(\{f_t\})$  and  $\mathcal{M}(\{f'_t\})$ , respectively. We say that  $\mathcal{M}$  is  $(\varepsilon, \delta)$ -DP if it is  $(\varepsilon, \delta)$ -DP at index  $i$  for all  $i \in [n]$ .

Notice that in light of the definition of  $\mathsf{E}_\gamma$ -divergence in (7), we can equivalently say  $\mathcal{M}$  is  $(\varepsilon, \delta)$ -DP if

$$\Pr(\mathcal{M}(\{f_t\}) \in A) - e^\varepsilon \Pr(\mathcal{M}(\{f'_t\}) \in A) \leq \delta,$$

for any  $A \subset \mathcal{W}$  and any pair of neighboring collections of cost functions  $\{f_t\}$  and  $\{f'_t\}$ . While this is a more widely used definition in DP literature, we will use its equivalent form, given in Definition 4, to emphasize the connection between  $\mathsf{E}_\gamma$ -divergence and DP.

**Remark 3.** *The above definition of DP is specific to online learning. Here, a pair of neighboring collections of cost functions is considered rather than the "neighboring datasets" used in offline learning scenarios. Definition 4 is also the standard notion of DP for online learning algorithm considered in [26–29, 107]. Nevertheless, Definition 4 differs from previous works on differentially private online algorithms in that privacy is ensured with respect to only the last parameter  $W_{n+1}$  rather than the entire set of parameters  $\{W_1, \dots, W_{n+1}\}$ .*

The analysis of DP mechanisms in online learning is particularly challenging: a single change in the algorithm's cost function at any time instance may have an accumulative impact on all future parameter updates. One standard way to address this issue is to add calibrated noise to  $\Psi_t$  at each iteration [26–29]. Thus, most differentially private online learning algorithms can be expressed as general iterative process of the form

$$W_{t+1} = \Pi_{\mathcal{W}}(\Psi_{f_t}(W_t) + \sigma_t Z_t), \quad (30)$$

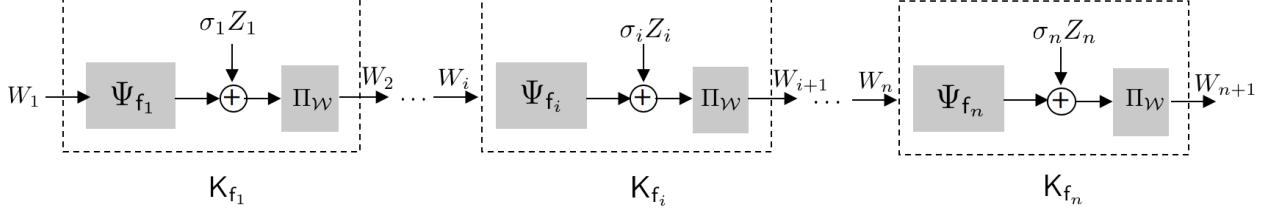


Fig. 2. The schematic representation of the iterative process described in (30). Each Markov kernel  $K_{f_t}$ ,  $t \in [n]$  consists of three components: mapping  $w \mapsto \Psi_{f_t}(w)$ , additive noise, and projection mapping  $\Pi_{\mathcal{W}}$ . The initial point  $W_1 \sim \mu_1$  is a random point in  $\mathcal{W}$ . Input and output of kernel  $K_{f_t}$  is  $W_t \sim \mu_t$  and  $W_{t+1} \sim \mu_{t+1}$ , respectively.

where  $\{Z_t\}$  is the collection of i.i.d. noise variables sampled from a known density with covariance matrix  $\mathbf{I}_d$  and  $\sigma_t$  specifies the magnitude of noise at time  $t$ .

The main goal of this section is to establish a bound for the DP parameters  $\varepsilon$  and  $\delta$  achievable by adding *Gaussian noise* to the update rules  $\{\Psi_{f_t}\}$ . The subsequent analysis is based on a key observation.

Assume  $\{Z_t\}$  are i.i.d. samples drawn from a Gaussian distribution. As illustrated in Fig. 2, the iterative process (30) can be viewed as a collection of Markov kernels  $\{K_{f_t}\}$  for  $t \in [n]$  where each  $K_{f_t}$  is a concatenation of the update rule  $\Psi_{f_t}$ , an additive Gaussian kernel with noise magnitude  $\sigma_t^2$ , and the projection operator  $\Pi_{\mathcal{W}}(\cdot)$ . Consequently, we can express  $K_{f_t}$  as  $\Pi_{\mathcal{W}} \circ K_t \circ \Psi_{f_t} : \mathcal{W} \rightarrow \mathcal{P}(\mathcal{W})$  where  $K_t$  is the  $\Psi_{f_t}(\mathcal{W})$ -constrained additive Gaussian kernel with noise magnitude given by  $\sigma_t^2$ . Note also that  $\mu_{t+1}$  the distribution of  $K_{f_t}$  is given by  $\mu_1 K_{f_1} \cdots K_{f_t}$ , where  $\mu_1$  is the distribution from which the initial point  $W_1$  is sampled. Similarly,  $\mu_{n+1}$  the distribution of the final parameter  $W_{n+1}$  is given by  $\mu_1 K_{f_1} \cdots K_{f_n}$ . If we change  $\{f_t\}$  to  $\{f'_t\}$  such that  $\{f_t\} \stackrel{i}{\sim} \{f'_t\}$ , then we denote the new distribution of the final parameter by  $\mu'_{n+1}$ . Since  $K_{f_t} = K_{f'_t}$  for all  $t \in [n] \setminus i$ , we can write

$$\begin{aligned} \mu_{n+1} &= \mu_i K_{f_i} K_{f_{i+1}} \cdots K_{f_n}, \\ \mu'_{n+1} &= \mu_i K_{f'_i} K_{f_{i+1}} \cdots K_{f_n}. \end{aligned}$$

A similar construction is given in [75] to quantify the improvement of DP parameters due to post-processing (aka privacy amplification). By repeatedly invoking the definition of contraction coefficient (4), we obtain

$$D_f(\mu_{n+1} \| \mu'_{n+1}) \leq D_f(\mu_i K_{f_i} \| \mu_i K_{f'_i}) \prod_{t=i+1}^n \eta_f(K_{f_t}),$$

and similarly for any  $\varepsilon \geq 0$

$$\mathbb{E}_{e^\varepsilon}(\mu_{n+1} \| \mu'_{n+1}) \leq \mathbb{E}_{e^\varepsilon}(\mu_i K_{f_i} \| \mu_i K_{f'_i}) \prod_{t=i+1}^n \eta_{e^\varepsilon}(K_{f_t}). \quad (31)$$

While this inequality follows from a routine application of SDPI, it provides a natural framework to study privacy guarantees of general iterative processes. In fact, inequality (31) indicates that an upper bound for DP parameters can be easily obtained by bounding  $\mathbb{E}_{e^\varepsilon}(\mu_i K_{f_i} \| \mu_i K_{f'_i})$  and  $\eta_{e^\varepsilon}(K_{f_i})$ . This is the linchpin for our privacy analysis in the sequel. Inequality (31) can also be used in applications beyond privacy. In particular, together with Theorem 2 and Proposition 1, it implies Theorem 1, which is of independent interest (e.g., in relation to studying information dissipation over Markov chain (1)).

**Theorem 5.** Assume that  $\mathcal{W} \subset \mathbb{R}^d$  is a closed convex set and  $P_Z = \mathcal{N}(0, \mathbf{I}_d)$ . The iterative process (30)

is  $(\varepsilon, \delta)$ -DP at index  $i \in [n]$  for  $\varepsilon \geq 0$  and

$$\delta = \theta_{e^\varepsilon} \left( \frac{\psi}{\sigma_i} \right) \prod_{t=i+1}^n \theta_{e^\varepsilon} \left( \frac{\text{dia}(\Psi_{f_t}(\mathcal{W}))}{\sigma_t} \right), \quad (32)$$

where  $\psi := \sup_{f_1, f_2 \in F} \sup_{w \in \mathcal{W}} \|\Psi_{f_1}(w) - \Psi_{f_2}(w)\|$ .

The detailed proof of this theorem is given in Appendix H. We will use this theorem to quantify the DP guarantee of popular algorithms. Before we delve into instantiating this result, we first “homogenize” Theorem 5 over all  $i \in [n]$  and give a bound for  $\delta$  independent of index  $i$ . To this end, we follow [73] to consider the *randomly-stopped* variant of process (30): Instead of terminating after pre-determined  $n$  iterations, the algorithm stops at a random time  $T$  uniformly chosen in  $[n]$ .

---

**Algorithm 1** Randomly-stopped variant of process (30)

---

**Require:** Collection of cost functions  $\{f_t\}$ , convex set  $\mathcal{W} \subset \mathbb{R}^d$ , and noise parameter  $\sigma$

- 1: Pick a starting point  $W_1 \sim \mu_1 \in \mathcal{P}(\mathcal{W})$  sampled from some distribution  $\mu_1$ ,
  - 2: Pick  $T$  uniformly at random from  $[n]$ ,
  - 3: **for**  $t \in \{1, \dots, T\}$  **do**
  - 4:   Play  $W_t$ , then obtain the cost function  $f_t$ ,
  - 5:    $W_{t+1} = \Pi_{\mathcal{W}}(\Psi_{f_t}(W_t) + \sigma_t Z_t)$ ,  $Z_t \sim \mathcal{N}(0, \mathbf{I}_d)$
  - 6: **end for**
  - 7: **return**  $W_{T+1}$ .
- 

**Theorem 6.** Assume that  $\mathcal{W} \subset \mathbb{R}^d$  is a closed convex set and  $P_Z = \mathcal{N}(0, \mathbf{I}_d)$ . The randomly-stopped iterative process described in Algorithm 1, is  $(\varepsilon, \delta)$ -DP for  $\varepsilon \geq 0$  and

$$\delta = \max_{i \in [n]} \left\{ \frac{1}{n} \sum_{t=i}^n \theta_{e^\varepsilon} \left( \frac{\psi}{\sigma_i} \right) \prod_{j=i+1}^t \theta_{e^\varepsilon} \left( \frac{\text{dia}(\Psi_{f_j}(\mathcal{W}))}{\sigma_j} \right) \right\} \quad (33)$$

$$\leq \frac{1}{n} \theta_{e^\varepsilon} \left( \frac{\psi}{\sigma} \right) \left[ 1 - \theta_{e^\varepsilon} \left( \frac{D}{\sigma} \right) \right]^{-1}, \quad (34)$$

where  $\psi := \sup_{f_1, f_2 \in F} \sup_{w \in \mathcal{W}} \|\Psi_{f_1}(w) - \Psi_{f_2}(w)\|$ ,  $\sigma := \min_{t \in [n]} \sigma_t$  and  $D := \max_{t \in [n]} \text{dia}(\Psi_{f_t}(\mathcal{W}))$ .

The detailed proof of this theorem is given in Appendix I. Similar to Theorem 5, the above result only requires  $\|\Psi_f(w)\|$  be uniformly bounded for all  $w \in \mathcal{W}$  and  $f \in F$ . This assumption is in fact weaker than the regularity conditions typically assumed in most differentially private iterative algorithms, such as [26, 77–88, 108, 109]. To illustrate this point and better compare our result with the existing differentially private ML algorithms, we instantiate Theorem 6 for two popular setups: (1) *one-pass* empirical risk minimization and (2) online gradient descent.

### B. Empirical Risk Minimization

Here, we consider an offline learning setting, namely empirical risk minimization (ERM) which is a fundamental problem in machine learning. Given a dataset  $\mathbb{D} = \{x_1, \dots, x_n\} \in \mathcal{X}^n$  and a loss function  $\ell : \mathcal{W} \times \mathcal{X} \rightarrow \mathbb{R}^+$ , ERM is formulated as

$$\inf_{w \in \mathcal{W}} \frac{1}{n} \sum_{t=1}^n \ell(w, x_t).$$

A simple algorithm for approximating the solution of this problem can be implemented with one pass of stochastic gradient descent (SGD) over the data set. This procedure starts with a random point  $W_1$  in  $\mathcal{W}$ , and then iterates  $W_{t+1} = \Pi_{\mathcal{W}}(W_t - \eta \nabla \ell(W_t, x_t))$  for  $t \in [n]$ . This problem can be cast as an online optimization problem by defining  $f_t(w) = \ell(w, x_t)$  and, thus, the differentially private version of such algorithm can be analyzed using Theorem 6. Following the DP literature (see, e.g., [78, 80, 81, 84, 88, 110, 111]), we add Gaussian noise to the gradient term, and thus the magnitude of the noise is assumed to be  $\eta_t^2 \sigma_t^2$  (as opposed to  $\sigma_t^2$ ). For brevity, we assume  $\eta_t = \eta$  and  $\sigma_t = \sigma$  for all  $t \in [n]$ . Consequently, we consider the following algorithm, which is known as *projected noisy SGD* (PNSGD)

$$W_{t+1} = \Pi_{\mathcal{W}}(W_t - \eta \nabla \ell(W_t, x_t) + \eta \sigma Z_t),$$

where  $Z_t \sim \mathcal{N}(0, \mathbf{I}_d)$ . It can be verified that the above algorithm is an instance of iterative process (30) with update function  $\Psi_{f_t}(w) = \Psi_{x_t}^{\text{SGD}}(w)$  where

$$\Psi_{x_t}^{\text{SGD}}(w) := w - \eta \nabla \ell(w, x_t).$$

In this setup each cost function  $f_t$  depends only on the data point  $x_t$ . Thus, the neighboring relationship between two collections of cost functions  $\{f_t\}$  and  $\{f'_t\}$  reduces to that of the datasets  $\mathbb{D} = \{x_1, \dots, x_n\}$  and  $\mathbb{D}' = \{x'_1, \dots, x'_n\}$ . Consequently, the definition of DP in Definition 4 can be equivalently given in terms of *neighboring datasets* and thus reduces to the original definition of DP in [112].

---

**Algorithm 2** Randomly stopped Projected Noisy SGD (PNSGD)

---

**Require:** Dataset  $\mathbb{D} = \{x_1, \dots, x_n\}$ , learning rate  $\eta > 0$ , convex set  $\mathcal{W} \subset \mathbb{R}^d$ , and noise parameter  $\sigma$

- 1: Pick  $W_1 \sim \mu_1 \in \mathcal{P}(\mathcal{W})$
- 2: Take  $T$  uniformly on  $[n]$
- 3: **for**  $t \in \{1, \dots, T\}$  **do**
- 4:      $W_{t+1} = \Pi_{\mathcal{W}}(W_t - \eta [\nabla_w \ell(W_t, x_t) + \sigma Z_t]), \quad Z_t \sim \mathcal{N}(0, \mathbf{I}_d)$
- 5: **end for**
- 6: **return**  $W_{T+1}$

---

The randomly-stopped PNSGD algorithm is described in Algorithm 2. The privacy guarantee of this algorithm has been recently studied by Feldman et al. [73] under the name of *privacy amplification by iteration*. However, the notion of privacy used in their work is Rényi differential privacy. In the following corollary, whose proof is given in Appendix J, we instantiate Theorem 6 to derive the privacy guarantee of this algorithm directly in terms of DP.

*Added in print:* After the first draft of this work, Sordello et al. [113] proposed another approach for homogenizing the privacy guarantee given in Theorem 5, specialized for PNSGD. Instead of randomizing the stopping time, they considered shuffling the dataset  $\mathbb{D}$  before initializing the algorithm and then applied our framework (in particular, Jensen's inequality and (31)). This indeed showcases the versatility of the contraction-based framework presented in our work.

**Corollary 5.** *Let  $\mathcal{W} \subset \mathbb{R}^d$  be a convex set and  $\{\ell(\cdot, x)\}_{x \in \mathcal{X}}$  be a family of convex  $L$ -Lipschitz functions over  $\mathcal{W}$ . Then the randomly-stopped PNSGD algorithm is  $(\varepsilon, \delta)$  with  $\varepsilon \geq 0$  and*

$$\delta = \frac{1}{n} \theta_{e^\varepsilon} \left( \frac{2L}{\sigma} \right) \left[ 1 - \theta_{e^\varepsilon} \left( \frac{\text{dia}(\mathcal{W}) + 2\eta L}{\eta \sigma} \right) \right]^{-1}. \quad (35)$$

If we further assume that  $w \mapsto \ell(w, x)$  is<sup>3</sup>  $\beta$ -smooth for any  $x \in \mathcal{X}$ , then a standard calculation in convex optimization shows that  $w \mapsto \Psi_x^{\text{SGD}}(w)$  is 1-Lipschitz for  $\eta \leq \frac{2}{\beta}$  (see Appendix K for a detailed

<sup>3</sup>A function  $f : \mathcal{W} \rightarrow \mathbb{R}$  is  $\beta$ -smooth if  $\|\nabla f(w_1) - \nabla f(w_2)\| \leq \beta \|w_1 - w_2\|$  for all  $w_1, w_2 \in \mathcal{W}$ .

proof). A similar argument as in the proof of Corollary 5 reveals that with this extra assumption (35) can be improved as (see Appendix K)

$$\delta = \frac{1}{n} \theta_{e^\varepsilon} \left( \frac{2L}{\sigma} \right) \left[ 1 - \theta_{e^\varepsilon} \left( \frac{\text{dia}(\mathcal{W})}{\eta\sigma} \right) \right]^{-1}, \quad (36)$$

for  $\eta \leq \frac{2}{\beta}$ . This enables us to formally compare our result with [73]. To do so, we need the following definition. Given  $\alpha > 1$ , a mechanism  $\mathcal{M}$  is called  $(\alpha, \zeta)$ -Rényi differentially-private (RDP) if

$$\sup_{\mathbb{D} \sim \mathbb{D}'} D_\alpha(\mu_{n+1} \| \mu'_{n+1}) \leq \zeta,$$

where  $D_\alpha(\cdot \| \cdot)$  denotes the Rényi divergence of order  $\alpha$  and  $\mu_{n+1}$  and  $\mu'_{n+1}$  are the distributions of  $W_{n+1}$  and  $W'_{n+1}$  the outputs of  $\mathcal{M}$  when running on neighboring  $\mathbb{D}$  and  $\mathbb{D}'$  (denoted by  $\mathbb{D} \sim \mathbb{D}'$ ), respectively.

**Theorem 7** ([73, Theorem 26]). *Let  $\mathcal{W} \subset \mathbb{R}^d$  be a convex set and  $\{\ell(\cdot, x)\}_{x \in \mathcal{X}}$  be a family of convex,  $L$ -Lipschitz and  $\beta$ -smooth loss functions over  $\mathcal{W}$ . Then, for any  $\eta \leq \frac{2}{\beta}$  and  $\alpha > 1$ , and  $\sigma \geq L\sqrt{2(\alpha-1)\alpha}$ , the randomly-stopped PNSGD algorithm is  $(\alpha, \zeta)$ -RDP for*

$$\zeta = \frac{4\alpha L^2 \log n}{n\sigma^2}.$$

While Corollary 5 provides the privacy guarantee for any  $\sigma > 0$ , this theorem is restricted to  $\sigma$  greater than  $L\sqrt{2(\alpha-1)\alpha}$ . This discrepancy stems from the fact that, unlike  $\mathsf{E}_\gamma$ -divergence,  $(\mu \| \nu) \rightarrow D_\alpha(\mu \| \nu)$  is not convex for  $\alpha > 1$ . To get around this issue, Feldman et al. [73, Lemma 25] presented a “weak” form of joint convexity for the Rényi divergence.

In order to compare Corollary 5 with Theorem 7, we need to convert  $(\alpha, \zeta)$ -RDP guarantee to  $(\varepsilon, \delta)$ -DP. To do so, we invoke two existing RDP-to-DP conversion formulae, making both analytical and numerical comparison possible.

**Lemma 6** ([90, Proposition 3]). *If a mechanism  $\mathcal{M}$  is  $(\alpha, \zeta)$ -RDP for  $\alpha > 1$ , then it is  $(\varepsilon, e^{-(\alpha-1)(\varepsilon-\zeta)})$ -DP.*

Due to the efficiency of RDP (especially in the context of composition), this lemma has been extensively used in many recent private ML algorithms, see e.g., [21, 110, 114, 115] and has been implemented in Google’s open-source TensorFlow Privacy<sup>4</sup> [117]. In light of this result, Theorem 7 implies that the randomly stopped PNSGD is  $(\varepsilon, \hat{\delta})$ -DP for any  $\varepsilon \geq 0$  and

$$\hat{\delta} := \inf_{\alpha \in (1, \alpha^*]} e^{-(\alpha-1)(\varepsilon-\zeta)}, \quad (37)$$

where  $\alpha^* := \frac{1}{2} \left[ 1 + \sqrt{1 + \frac{2\sigma^2}{L^2}} \right]$  and  $\zeta = \frac{4\alpha L^2 \log n}{n\sigma^2}$ . The restriction on the range of  $\alpha$  was caused by the constraint on  $\sigma$  in Theorem 7. Since  $\zeta$  is linear in  $\alpha$ , the minimization in (37) can be solved analytically:

$$\hat{\delta} = e^{-(\alpha^{\text{opt}}-1)(\varepsilon-\rho\alpha^{\text{opt}})}, \quad (38)$$

where  $\rho := \frac{4L^2 \log n}{n\sigma^2}$  and

$$\alpha^{\text{opt}} := \begin{cases} \alpha^*, & \text{if } \varepsilon^2 \geq \rho^2 \left( 1 + \frac{2\sigma^2}{L^2} \right), \\ \frac{1}{2} + \frac{\varepsilon}{2\rho}, & \text{otherwise.} \end{cases}$$

Hence, for sufficiently large  $\varepsilon$  (i.e.,  $\varepsilon \geq O(\frac{\log n}{n})$ ),  $\hat{\delta}$  behaves approximately like  $e^{-O(\varepsilon)}$  while our privacy analysis yields  $\delta$  in (36) that decays as  $e^{-O(\varepsilon^2)}$ .

<sup>4</sup>See the function “compute\_eps” in the analysis directory of TensorFlow Privacy in [116]. At the time of submission of this paper, Lemma 6 is particularly implemented in the line 216 of [116].

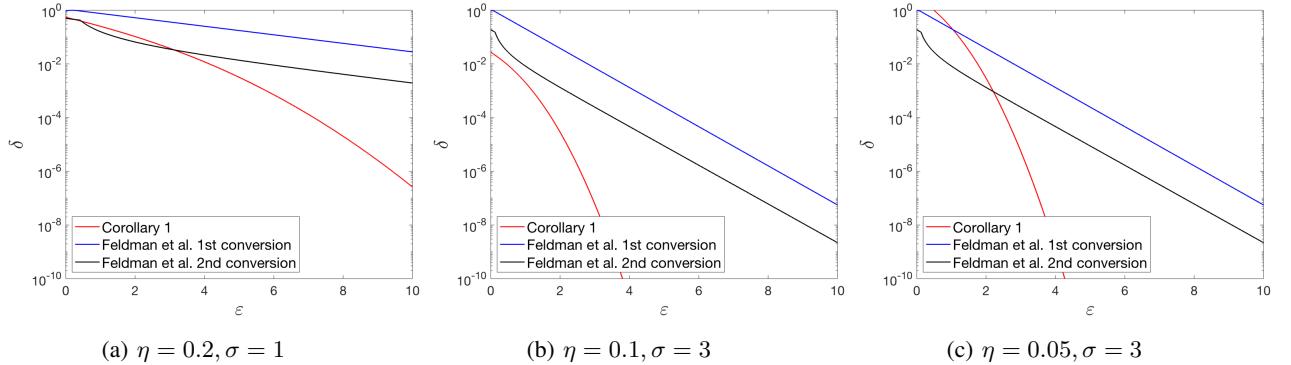


Fig. 3. The privacy parameters of PNSGD Algorithm 2 obtained from Corollary 5 and converting [73, Theorem 26] according to (38) and (39), respectively. Here, we vary the learning rate  $\eta$  and  $\sigma$ . Other parameters are as follows:  $L = 1$ ,  $\beta = 1$ , and  $n = 100$ .

Despite its prevalence in practice, Lemma 6 is not tight in general. This issue has recently prompted Assoodeh et al. [118] to derive the *optimal* translation of  $(\alpha, \zeta)$ -RDP into  $(\varepsilon, \delta)$ -DP. Invoking [118, Lemma 2], we deduce that the randomly stopped PNSGD is  $(\varepsilon, \check{\delta})$ -DP for  $\varepsilon \geq 0$  and

$$\check{\delta} := \inf_{\alpha \in (1, \alpha^*]} \min \left\{ \kappa e^{-(\alpha-1)(\varepsilon-\zeta)}, \frac{e^{(\alpha-1)\zeta} - 1}{\alpha(e^{(\alpha-1)\varepsilon} - 1)} \right\}, \quad (39)$$

where  $\kappa := \frac{1}{\alpha} \left(1 - \frac{1}{\alpha}\right)^{\alpha-1}$  and  $\zeta = \frac{4\alpha L^2 \log n}{n\sigma^2}$ . Although this result is tighter than Lemma 6, it is not possible to analytically express  $\check{\delta}$ . In Fig. 3, we compare the privacy parameters of the randomly-stopped PNSGD given in Corollary 5 with  $\hat{\delta}$  and  $\check{\delta}$  given in (38) and (39), respectively. As illustrated in this figure, Corollary 5 results in significantly smaller values of  $\delta$ , compared to the technique developed in [73], for sufficiently large  $\varepsilon$  across different values of parameters  $\eta$  and  $\sigma$ . The privacy improvement of Corollary 5 over [73] is more pronounced for smaller values of  $\frac{\eta}{\sigma}$ . In particular, for the reasonable range<sup>5</sup> of  $\eta \in (0.05, 0.1)$  our technique results in tighter privacy parameters than [73] for all  $\varepsilon \geq 2$  as long as  $\sigma \geq 3$ . It is also worth emphasizing that, unlike [73], the privacy parameter given in Corollary 5 heavily depends on the learning rate  $\eta$ . Intuitively, the larger the value of  $\eta$ , the stronger the privacy guarantee of the PNSGD algorithm. Corollary 5 formalizes this intuition by establishing a precise rate at which  $\delta$  decreases as  $\eta$  increases for any given  $\varepsilon$ .

### C. Online Gradient Descent Algorithm

In this section, we apply Theorem 6 to study the privacy guarantee of the online gradient descent (OGD) algorithm [59]. Then, by drawing on standard results from online convex optimization, we expound the trade-off between privacy and utility achievable by this algorithm.

The OGD algorithm is an instance of the iterative process (30) with update function

$$\Psi_{f_t}^{\text{OGD}}(w) := w - \eta_t \nabla f_t(w).$$

Similar to Section IV-B, we add Gaussian noise to the gradient term and thus the noisy OGD algorithm iterates as

$$W_{t+1} = \Pi_{\mathcal{W}}(W_t - \eta_t \nabla f_t(W_t) + \eta_t \sigma_t Z_t), \quad (40)$$

<sup>5</sup>Abadi et al. [110] empirically observed that the accuracy of PNSGD algorithm is stable for a learning rate in the range of (0.01, 0.07) and peaks around 0.052.

where  $Z_t \sim \mathcal{N}(0, \mathbf{I}_d)$ . A standard quantity for measuring the utility of typical online algorithms is regret described in Section IV-C. To take into account our assumptions that the stopping time is random and *only*  $W_{n+1}$  is available for analysis, we choose to measure utility in terms of *stochastic regret* [119]. Recall that  $\mathcal{F}$  denotes the collection of all possible real-valued cost functions over  $\mathcal{W}$ . Here, we assume that  $\mathcal{F}$  consists of all convex functions over  $\mathcal{W}$  with uniformly bounded gradients and let  $\mu$  be a distribution over  $\mathcal{F}$ . We further assume that the cost functions  $\{\mathbf{f}_t\}$  are independently sampled from  $\mathcal{F}$  according to  $\mu$ . The *stochastic regret* for the randomly stopped OGD algorithm is given by

$$SR(n) := \mathbb{E}[\mathbf{f}(W_T)] - \inf_{w \in \mathcal{W}} \mathbf{f}(w),$$

where the expectation is taken with respect to the randomness in stopping time  $T$  and  $\mathbf{f}(w) = \mathbb{E}_{\mu}[\mathbf{f}_t(w)]$ . It is worth noting that stochastic regret is sometimes referred to as *expected excess population loss* in the offline setting, see e.g., [74, 81, 120, 121].

**Proposition 2.** *Assume that  $\mathcal{W} \subset \mathbb{R}^d$  is a closed convex set,  $\mathcal{F}$  is a family of convex functions over  $\mathcal{W}$  with uniformly bounded gradients,  $\{\mathbf{f}_t\}$  are  $n$  independent samples from  $\mathcal{F}$ , and  $P_Z = \mathcal{N}(0, \mathbf{I}_d)$ . If  $\eta_t = \frac{\text{dia}(\mathcal{W})}{M\sqrt{t}}$  with  $M = \sup_{\mathbf{f}_1 \in \mathcal{F}} \sup_{w \in \mathcal{W}} \|\nabla \mathbf{f}_1(w)\|$ , then the randomly stopped OGD algorithm satisfies*

$$SR(n) \leq \frac{3M\text{dia}(\mathcal{W})}{2\sqrt{n}} + \frac{d}{2n} \sum_{t=1}^n \eta_t \sigma_t^2.$$

This proposition follows from standard results in online convex optimization, see, e.g., [49]. For the reader's convenience, we provide a full proof in Appendix L. Combined with Theorem 6, this proposition can be applied to capture the balance between privacy and utility of OGD algorithm. In particular, they elucidate that if  $\eta_t = O(\frac{1}{\sqrt{t}})$ , then a necessary condition for non-trivial privacy and utility is  $\sigma_t \rightarrow \infty$  and  $\frac{\sigma_t}{\sqrt{t}} \rightarrow 0$ , i.e.,  $\{\sigma_t\}$  must be an increasing sequence but with a rate slower than  $\sqrt{t}$ .

It is worth noting that Feldman et al. [74] recently proposed an offline algorithm, termed *Snowball-SGD*, which is shown to be order optimal in both the expected excess population loss and also privacy guarantee. However, this algorithm differs from ours most notably in the batch size. While the batch size in our analysis is assumed to be one and the stopping time is random, in Snowball-SGD the batch size gradually increases from one to  $\lceil \sqrt{d} \rceil$  as the algorithm progresses.

## REFERENCES

- [1] S. Asoodeh, M. Diaz, and F. P. Calmon, "Privacy amplification of iterative algorithms via contraction coefficients," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2020.
- [2] S. Asoodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an  $f$ -divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 545–550.
- [3] R. Ahlsweide and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, 12 1976.
- [4] R. L. Dobrushin, "Central limit theorem for nonstationary markov chains. I," *Theory Probab. Appl.*, vol. 1, no. 1, pp. 65–80, 1956.
- [5] A. Kontorovich and M. Raginsky, "Concentration of measure without independence: A unified approach via the martingale method," in *Convexity and Concentration*. Springer New York, 2017, pp. 183–210.
- [6] A. Makur and L. Zheng, "Comparison of contraction coefficients for  $f$ -divergences," *Probl. Inf. Trans.*, vol. 56, pp. 103–156, 2020.
- [7] M. Raginsky, "Strong data processing inequalities and  $\phi$ -sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.
- [8] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, Jan 2016.

- [9] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*, E. Carlen, M. Madiman, and E. M. Werner, Eds. New York, NY: Springer New York, 2017, pp. 211–249.
- [10] F. d. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1879–1892, 2018.
- [11] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of Royal Statistics*, vol. 28, pp. 131–142, 1966.
- [12] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
- [13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
- [14] N. Sharma and N. A. Warsi, "Fundamental bound on the reliability of quantum information transmission," *CoRR*, vol. abs/1302.5281, 2013. [Online]. Available: <http://arxiv.org/abs/1302.5281>
- [15] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Commun. Inf. Theory*, vol. 1, no. 4, pp. 417–528, Dec. 2004.
- [16] J. Cohen, J. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices, with Applications in Information Theory, Statistics, Economics, and Population Sciences*. Birkhäuser, 1998.
- [17] R. Subramanian, B. N. Vellambi, and I. Land, "An improved bound on information loss due to finite block length in a gaussian line network," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 1864–1868.
- [18] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2015, pp. 2376–2380.
- [19] L. LeCam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, no. 1, pp. 38–53, 01 1973. [Online]. Available: <https://doi.org/10.1214/aos/1193342380>
- [20] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy, data processing inequalities, and statistical minimax rates," in *Proc. Symp. Foundations of Computer Science*, 2013, p. 429–438.
- [21] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018.
- [22] J. Duchi and R. Rogers, "Lower bounds for locally private estimation via communication complexity," in *Proc. of the Thirty-Second Conference on Learning Theory (COLT)*, vol. 99, 25–28 Jun 2019, pp. 1161–1191.
- [23] A. Rohde and L. Steinberger, "Geometrizing rates of convergence under local differential privacy constraints," *The Annals of Statistics*, vol. 48, no. 5, pp. 2646 – 2670, 2020.
- [24] J. Acharya and Z. Sun, "Communication complexity in locally private distribution estimation and heavy hitters," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 51–60. [Online]. Available: <https://proceedings.mlr.press/v97/acharya19c.html>
- [25] J. Acharya, C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi, "Inference under information constraints iii: Local privacy constraints," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 253–267, 2021.
- [26] P. Jain, P. Kothari, and A. Thakurta, "Differentially private online learning," in *Proc. Conference on Learning Theory (COLT)*, vol. 23, 25–27 Jun 2012, pp. 24.1–24.34.
- [27] A. Guha Thakurta and A. Smith, "(Nearly) optimal algorithms for private online learning in full-information and bandit settings," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 2733–2741.
- [28] N. Agarwal and K. Singh, "The price of differential privacy for online learning," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70. PMLR, 06–11 Aug 2017, pp. 32–40.
- [29] A. R. Cardoso and R. Cummings, "Differentially private online submodular minimization," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 89, 16–18 Apr 2019, pp. 1650–1658.
- [30] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *2014 IEEE Int. Symp. Inf. Theory*, 2014, pp. 3022–3026.
- [31] A. Makur and L. Zheng, "Bounds between contraction coefficients," 2018. [Online]. Available: <http://arxiv.org/abs/1510.01844>
- [32] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*, E. Carlen, M. Madiman, and E. M. Werner, Eds. New York, NY: Springer

- New York, 2017, pp. 211–249.
- [33] A. Makur and Y. Polyanskiy, “Comparison of channels: Criteria for domination by a symmetric channel,” *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5704–5725, 2018.
  - [34] A. Xu and M. Raginsky, “Converses for distributed estimation via strong data processing inequalities,” in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 2376–2380.
  - [35] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, “Communication lower bounds for statistical estimation problems via a distributed data processing inequality,” in *Proc. of the Forty-Eighth Annual ACM Symposium on Theory of Computing (STOC)*, 2016, p. 1011–1020.
  - [36] A. Xu and M. Raginsky, “Information-theoretic lower bounds for distributed function computation,” *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2314–2337, 2017.
  - [37] M. Joseph, J. Kulkarni, J. Mao, and S. Z. Wu, “Locally private gaussian estimation,” in *Advances in Neural Information Processing Systems 32*, 2019, pp. 2984–2993.
  - [38] P. Kamalaruban, “Transitions, losses, and re-parameterizations: Elements of prediction games,” Ph.D. dissertation, The Australian National University, 2018.
  - [39] M. H. DeGroot, “Uncertainty, information, and sequential experiments,” *Ann. Math. Statist.*, vol. 33, no. 2, pp. 404–419, 06 1962.
  - [40] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, “What can we learn privately?” *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, Jun. 2011.
  - [41] M. Gaboardi, R. Rogers, and O. Sheffet, “Locally private mean estimation:  $z$ -test and tight confidence intervals,” in *Proc. Machine Learning Research*, 2019, pp. 2545–2554.
  - [42] P. Kairouz, S. Oh, and P. Viswanath, “Extremal mechanisms for local differential privacy,” *Journal of Machine Learning Research*, vol. 17, no. 17, pp. 1–51, 2016.
  - [43] L. P. Barnes, W. N. Chen, and A. Özgür, “Fisher information under local differential privacy,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 645–659, 2020.
  - [44] J. Acharya, C. L. Canonne, and H. Tyagi, “Inference under information constraints i: Lower bounds from chi-square contraction,” *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7835–7855, 2020.
  - [45] M. Ye and A. Barg, “Optimal schemes for discrete distribution estimation under locally differential privacy,” *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5662–5676, 2018.
  - [46] D. Wang and J. Xu, “On sparse linear regression in the local differential privacy model,” *IEEE Trans. Inf. Theory*, pp. 1–1, 2020.
  - [47] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting privacy breaches in privacy preserving data mining,” in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2003, pp. 211–222.
  - [48] T. Berrett and C. Butucea, “Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3164–3173.
  - [49] E. Hazan, “Introduction to online convex optimization,” *Found. Trends Optim.*, vol. 2, no. 3–4, p. 157–325, Aug. 2016.
  - [50] S. Shalev-Shwartz, “Online learning and online convex optimization,” *Found. Trends Mach. Learn.*, vol. 4, no. 2, p. 107–194, Feb. 2012.
  - [51] E. Takimoto and M. K. Warmuth, “Path kernels and multiplicative updates,” *Journal of Machine Learning Research*, vol. 4, pp. 773–818, Oct. 2003.
  - [52] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Identifying suspicious urls: An application of large-scale online learning,” in *Proc. International Conference on Machine Learning*, ser. ICML ’09. New York, NY, USA: Association for Computing Machinery, 2009, p. 681–688.
  - [53] T. M. Cover, “Universal portfolios,” *Mathematical Finance*, vol. 1, no. 1, pp. 1–29, 1991.
  - [54] B. Li, S. Hoi, P. Zhao, and V. Gopalkrishnan, “Confidence weighted mean reversion strategy for on-line portfolio selection,” in *International Conference on Artificial Intelligence and Statistics (AISTAT)*, vol. 15, 2011, pp. 434–442.
  - [55] A. Kalai and S. Vempala, “Efficient algorithms for universal portfolios,” *Journal of Machine Learning Research*, vol. 3, p. 423–440, 2002.
  - [56] A. Ben-Tal, E. Hazan, T. Koren, and S. Mannor, “Oracle-based robust optimization via online learning,” *Operations Research*, vol. 63, no. 3, pp. 628–638, June 2015.
  - [57] E. Hazan and S. Kale, “An optimal algorithm for stochastic strongly-convex optimization,” *Journal of Machine*

- Learning Research*, vol. 15, p. 2489– 2512, 2014.
- [58] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, “Differential privacy under continual observation,” in *Proc. of the Forty-Second ACM Symposium on Theory of Computing*, ser. STOC ’10, 2010, p. 715–724.
- [59] M. Zinkevich, “Online convex programming and generalized infinitesimal gradient ascent,” in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML)*, 2003, p. 928–935.
- [60] B. Kulis and P. L. Bartlett, “Implicit online learning,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, 2010, pp. 575–582.
- [61] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, “Online passive-aggressive algorithms,” *J. Mach. Learn. Res.*, vol. 7, p. 551–585, Dec. 2006.
- [62] J. C. Duchi, S. Shalev-shwartz, Y. Singer, and A. Tewari, “Composite objective mirror descent,” in *Proc. Conference on Learning Theory (COLT)*, 2010, pp. 14–26.
- [63] M. Frank and P. Wolfe, “An algorithm for quadratic programming,” *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.
- [64] J. Abernethy, E. Hazan, and A. Rakhlin, “Competing in the dark: An efficient algorithm for bandit linear optimization,” 2008. [Online]. Available: <http://www.mit.edu/~rakhlin/papers/AbeHazRak08.pdf>
- [65] S. Shalev-Shwartz and Y. Singer, “A primal-dual perspective of online learning algorithms,” *Mach. Learn.*, vol. 69, no. 2–3, p. 115–142, Dec. 2007.
- [66] E. Hazan, “Extracting certainty from uncertainty: Regret bounded by variation in costs,” in *In COLT*, 2008.
- [67] B. McMahan, “Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization,” in *Proc. International Conference on Artificial Intelligence and Statistics (AISTAT)*, vol. 15, 11–13 Apr 2011, pp. 525–533.
- [68] A. Rakhlin, O. Shamir, and K. Sridharan, “Making gradient descent optimal for strongly convex stochastic optimization,” in *Proc. International Conference on International Conference on Machine Learning*, 2012, p. 1571–1578.
- [69] O. Shamir and T. Zhang, “Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes,” in *Proc. International Conference on Machine Learning*, vol. 28, no. 1. PMLR, 17–19 Jun 2013, pp. 71–79.
- [70] J. Abernethy, K. A. Lai, and A. Wibisono, “Last-iterate convergence rates for min-max optimization,” 2019.
- [71] Q. Lei, S. G. Nagarajan, I. Panageas, and X. Wang, “Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes,” 2020.
- [72] Y. Lei, P. Yang, K. Tang, and D.-X. Zhou, “Optimal stochastic and online learning with individual iterates,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 5415–5425.
- [73] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, “Privacy amplification by iteration,” *FOCS*, pp. 521–532, 2018.
- [74] V. Feldman, T. Koren, and K. Talwar, “Private stochastic convex optimization: Optimal rates in linear time,” in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2020, 2020, p. 439–449.
- [75] B. Balle, G. Barthe, M. Gaboardi, and J. Geumlek, “Privacy amplification by mixing and diffusion mechanisms,” in *NeurIPS*, 2019, pp. 13 277–13 287.
- [76] S. Augenstein, H. B. McMahan, D. Ramage, S. Ramaswamy, P. Kairouz, M. Chen, R. Mathews, and B. A. y Arcas, “Generative models for effective ML on private, decentralized datasets,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=SJgaRA4FPH>
- [77] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton, “Bolt-on differential privacy for scalable stochastic gradient descent-based analytics,” in *SIGMOD*, 2017, pp. 1307–1322.
- [78] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, “Differentially private empirical risk minimization,” *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.
- [79] K. Chaudhuri and N. Mishra, “When random sampling preserves privacy,” in *Advances in Cryptology - CRYPTO 2006*. Springer Berlin Heidelberg, 2006, pp. 198–213.
- [80] R. Bassily, A. Smith, and A. Thakurta, “Private empirical risk minimization, revisited,” in *ICML 2014 Workshop on Learning, Security and Privacy*, Beijing, China, 25 Jun 2014.
- [81] R. Bassily, V. Feldman, K. Talwar, and A. Guha Thakurta, “Private stochastic convex optimization with optimal rates,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 11 282–11 291.
- [82] K. Chaudhuri and C. Monteleoni, “Privacy-preserving logistic regression,” in *Advances in Neural Information*

- Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2009, pp. 289–296.
- [83] A. G. Thakurta and A. Smith, “Differentially private feature selection via stability arguments, and the robustness of the lasso,” in *Proceedings of the 26th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, S. Shalev-Shwartz and I. Steinwart, Eds., vol. 30. Princeton, NJ, USA: PMLR, 12–14 Jun 2013, pp. 819–850.
- [84] S. Song, K. Chaudhuri, and A. D. Sarwate, “Stochastic gradient descent with differentially private updates,” in *2013 IEEE Global Conference on Signal and Information Processing*, 2013, pp. 245–248.
- [85] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in *Proc. of IEEE Foundations of Computer Science (FOCS)*, 2013.
- [86] A. Smith, A. Thakurta, and J. Upadhyay, “Is interaction necessary for distributed private learning?” in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 58–77.
- [87] K. Talwar, A. Thakurta, and L. Zhang, “Nearly-optimal private lasso,” in *Proc. International Conference on Neural Information Processing Systems (NeurIPS)*, Cambridge, MA, USA, 2015, p. 3025–3033.
- [88] D. Wang, M. Ye, and J. Xu, “Differentially private empirical risk minimization revisited: Faster and more general,” in *Proc. of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, p. 2719–2728.
- [89] S. M. Kakade and A. Tewari, “On the generalization ability of online strongly convex programming algorithms,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2009, pp. 801–808.
- [90] I. Mironov, “Rényi differential privacy,” in *Proc. Computer Security Found. (CSF)*, 2017, pp. 263–275.
- [91] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, “Three variants of differential privacy: Lossless conversion and applications,” *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 208–222, 2021.
- [92] P. Billingsley, *Probability and Measure*, 3rd ed. John Wiley and Sons, 1995.
- [93] J. Liu, P. Cuff, and S. Verdú, “ $E_\gamma$ -resolvability,” *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 2629–2658, May 2017.
- [94] P. Del Moral, M. Ledoux, and L. Miclo, “On contraction properties of markov kernels,” *Probab. Theory Relat. Fields*, vol. 126, pp. 395–420, 2003.
- [95] B. Balle and Y.-X. Wang, “Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 394–403.
- [96] I. Sason and S. Verdú, “ $f$ -divergence inequalities,” *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.
- [97] S. L. Warner, “Randomized response: A survey technique for eliminating evasive answer bias,” *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.
- [98] P. Kairouz, K. Bonawitz, and D. Ramage, “Discrete distribution estimation under local privacy,” in *Proc. Int. Conf. Machine Learning*, vol. 48, 20–22 Jun 2016, pp. 2436–2444.
- [99] B. Yu, *Assouad, Fano, and Le Cam*. Springer New York, 1997, pp. 423–435.
- [100] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 10 1999.
- [101] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008.
- [102] J. Bretagnolle and C. Huber, “Estimation des densités: risque minimax,” *Séminaire de Probabilités XII*, pp. 342–363, 1978.
- [103] I. Vajda, “Note on discrimination information and variation (corresp.),” *IEEE Transactions on Information Theory*, vol. 16, no. 6, pp. 771–773, 1970.
- [104] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [105] F. McSherry and I. Mironov, “Differentially private recommender systems: Building privacy into the netflix prize contenders,” in *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009, p. 627–636.
- [106] R. Guerraoui, A.-M. Kermarrec, R. Patra, and M. Taziki, “D2p: Distance-based differential privacy in recommenders,” *Proc. VLDB Endow.*, vol. 8, no. 8, p. 862–873, Apr. 2015.
- [107] E. Nozari, P. Tallapragada, and J. Cortés, “Differentially private distributed convex optimization via functional perturbation,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 395–408, March 2018.
- [108] P. Jain and A. G. Thakurta, “(near) dimension independent risk bounds for differentially private learning,” in *Proc. of the 31st International Conference on Machine Learning (ICML)*, vol. 32, no. 1, 22–24 Jun 2014,

- pp. 476–484.
- [109] K. S. S. Kumar and M. P. Deisenroth, “Differentially private empirical risk minimization with sparsity-inducing norms,” *arXiv:1905.04873*, 2019.
  - [110] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proc. ACM SIGSAC CCS*, 2016, pp. 308–318.
  - [111] R. Bassily, A. Smith, and A. Thakurta, “Private empirical risk minimization: Efficient algorithms and tight error bounds,” in *Proc. of IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, 2014, pp. 464–473.
  - [112] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Proc. Theory of Cryptography (TCC)*, Berlin, Heidelberg, 2006, pp. 265–284.
  - [113] M. Sordello, Z. Bu, and J. Dong, “Privacy amplification via iteration for shuffled and online PNSGD,” in *Machine Learning and Knowledge Discovery in Databases. Research Track*, N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano, Eds., 2021, pp. 796–813.
  - [114] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, “Subsampled Rényi differential privacy and analytical moments accountant,” in *AISTAT*, vol. 89, 16–18 Apr 2018, pp. 1226–1235.
  - [115] B. McMahan, G. Andrew, I. Mironov, N. Papernot, P. Kairouz, S. Chien, and U. Erlingsson, “A general approach to adding differential privacy to iterative training procedures,” in *Workshop on Privacy Preserving Machine Learning (NeurIPS 2018)*, 2018. [Online]. Available: <https://arxiv.org/pdf/1812.06210.pdf>
  - [116] [https://github.com/tensorflow/privacy/blob/master/tensorflow\\_privacy/privacy/analysis/rdp\\_accountant.py](https://github.com/tensorflow/privacy/blob/master/tensorflow_privacy/privacy/analysis/rdp_accountant.py), [Online; accessed 15-December-2020].
  - [117] TensorFlow Privacy, 2019. [Online]. Available: <https://github.com/tensorflow/privacy>
  - [118] S. Assoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, “A better bound gives a hundred rounds: Enhanced privacy guarantees via  $f$ -divergences,” in *Proc. of Int. Symp. Inf Theory (ISIT)*, 2020. [Online]. Available: <https://arxiv.org/abs/2001.05990>
  - [119] L. Chen, C. Harshaw, H. Hassani, and A. Karbasi, “Projection-free online optimization with stochastic gradient: From convexity to submodularity,” in *Proc. of the 35th International Conference on Machine Learning (ICML)*, vol. 80, 10–15 Jul 2018, pp. 814–823.
  - [120] V. Feldman and J. Vondrak, “High probability generalization bounds for uniformly stable algorithms with nearly optimal rate,” in *Conference on Learning Theory*, 2019, pp. 1270–1279.
  - [121] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar, “Stability of stochastic gradient descent on nonsmooth convex losses,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 4381–4391.
  - [122] S. Bubeck, *Convex Optimization: Algorithms and Complexity*. Foundations and Trends in Machine Learning, 2015, vol. 8, no. 3–4.

## APPENDIX

### A. Proof of Theorem 2

Observe that the case  $\gamma = 1$  corresponds to Dobrushin’s formula (5), so we assume that  $\gamma > 1$ .

We begin by showing that, for any two probability measures  $\mu, \nu \in \mathcal{P}(\mathcal{D})$ ,

$$\mathsf{E}_\gamma(\mu \mathsf{K} \| \nu \mathsf{K}) \leq \mathsf{E}_\gamma(\mu \| \nu) \sup_{x_1, x_2 \in \mathcal{D}} \mathsf{E}_\gamma(\mathsf{K}(\cdot | x_1) \| \mathsf{K}(\cdot | x_2)). \quad (41)$$

Let  $\phi := \mu - \gamma\nu$  and  $(\phi^+, \phi^-)$  be its Hahn-Jordan decomposition. By the definition of  $\mathsf{E}_\gamma$  in (6), we have that  $\mathsf{E}_\gamma(\mu \| \nu) = \|\phi^+\|$  and

$$\begin{aligned} \mathsf{E}_\gamma(\mu \| \nu) &= \frac{1}{2}\|\phi^+\| + \frac{1}{2}\|\phi^-\| + \frac{1}{2}\|\phi^+\| - \frac{1}{2}\|\phi^-\| \\ &= \frac{1}{2}\|\phi\| + \frac{1}{2}(1 - \gamma), \end{aligned} \quad (42)$$

where the last equality follows from the fact that  $\phi^+$  and  $\phi^-$  are positive measures and thus

$$\|\phi^+\| - \|\phi^-\| = \phi^+(\mathcal{D}) - \phi^-(\mathcal{D}) = \phi(\mathcal{D}) = 1 - \gamma. \quad (43)$$

*Mutatis mutandis*, we have that

$$\mathbb{E}_\gamma(\mu K\| \nu K) = \|(\phi K)^+\| = \frac{1}{2}\|\phi K\| + \frac{1}{2}(1-\gamma). \quad (44)$$

It follows from the definition of the Hahn-Jordan decomposition that if  $\phi^+ \equiv 0$ , then  $(\phi K)^+ \equiv 0$ . In this case, (41) holds trivially as

$$\mathbb{E}_\gamma(\mu K\| \nu K) = \|(\phi K)^+\| = 0 = \|\phi^+\| = \mathbb{E}_\gamma(\mu\|\nu).$$

Now assume that  $\phi^+$  is not the trivial measure. By (43),

$$\phi^+(\mathcal{D}) - \phi^-(\mathcal{D}) = 1 - \gamma < 0,$$

which implies that  $\phi^-(\mathcal{D}) > 0$ , i.e.,  $\phi^-$  is not the trivial measure. Hence, both  $\phi^+/\|\phi^+\|$  and  $\phi^-/\|\phi^-\|$  are well-defined probability measures. As a result, we have that

$$\begin{aligned} \|\phi K\| &= \left\| \int K(\cdot|x_1) d\phi^+(x_1) - \int K(\cdot|x_2) d\phi^-(x_2) \right\| \\ &= \left\| \int \int [\|\phi^+\|K(\cdot|x_1) - \|\phi^-\|K(\cdot|x_2)] \frac{d\phi^+(x_1)}{\|\phi^+\|} \frac{d\phi^-(x_2)}{\|\phi^-\|} \right\| \\ &\leq \sup_{x_1, x_2 \in \mathcal{D}} \|\|\phi^+\|K(\cdot|x_1) - \|\phi^-\|K(\cdot|x_2)\|. \end{aligned} \quad (45)$$

For ease of notation, let

$$\psi(x_1, x_2) := \|\|\phi^+\|K(\cdot|x_1) - \|\phi^-\|K(\cdot|x_2)\|.$$

By adding and subtracting the term  $\gamma\|\phi^+\|K(\cdot|x_2)$ , the triangle inequality implies that

$$\psi(x_1, x_2) \leq \|\phi^+\|\|K(\cdot|x_1) - \gamma K(\cdot|x_2)\| + \|\phi^-\| - \gamma\|\phi^+\|,$$

where we used the inequality  $\|\phi^-\| - \gamma\|\phi^+\| \geq 0$ . In addition, a direct computation shows that

$$\psi(x_1, x_2) \leq \|\phi^+\| (\|K(\cdot|x_1) - \gamma K(\cdot|x_2)\| + 1 - \gamma) - (\|\phi^+\| - \|\phi^-\|).$$

Therefore, (42) and (43) imply that

$$\psi(x_1, x_2) \leq 2\|\phi^+\|\mathbb{E}_\gamma(K(\cdot|x_1)\|K(\cdot|x_2)) - (1 - \gamma).$$

As a result, (45) becomes

$$\|\phi K\| \leq 2\|\phi^+\| \sup_{x_1, x_2 \in \mathcal{D}} \mathbb{E}_\gamma(K(\cdot|x_1)\|K(\cdot|x_2)) - (1 - \gamma).$$

By (44) and the fact that  $\mathbb{E}_\gamma(\mu\|\nu) = \|\phi^+\|$ , the previous inequality becomes

$$\mathbb{E}_\gamma(\mu K\| \nu K) \leq \mathbb{E}_\gamma(\mu\|\nu) \sup_{x_1, x_2 \in \mathcal{D}} \mathbb{E}_\gamma(K(\cdot|x_1)\|K(\cdot|x_2))$$

or, equivalently,

$$\eta_\gamma(K) \leq \sup_{x_1, x_2 \in \mathcal{D}} \mathbb{E}_\gamma(K(\cdot|x_1)\|K(\cdot|x_2)). \quad (46)$$

Now we show that the reverse inequality in (46) holds, and hence the desired equality. For  $x \in \mathcal{D}$ , let  $\delta_x$  be the Dirac mass at  $x$ . Observe that for  $x_1, x_2 \in \mathcal{D}$  such that  $x_1 \neq x_2$ , we have that

$$\mathbb{E}_\gamma(\delta_{x_1}\|\delta_{x_2}) = (\delta_{x_1} - \gamma\delta_{x_2})^+(\mathcal{D}) = 1.$$

Since  $\delta_x K = K(\cdot|x)$  for every  $x \in \mathcal{D}$ , the previous equality implies that

$$\frac{E_\gamma(\delta_{x_1} K \| \delta_{x_2} K)}{E_\gamma(\delta_{x_1} \| \delta_{x_2})} = E_\gamma(K(\cdot|x_1) \| K(\cdot|x_2)).$$

Therefore, by the definition of  $\eta_\gamma(K)$  in (4),

$$\begin{aligned}\eta_\gamma(K) &\geq \sup_{x_1 \neq x_2} \frac{E_\gamma(\delta_{x_1} K \| \delta_{x_2} K)}{E_\gamma(\delta_{x_1} \| \delta_{x_2})} \\ &= \sup_{x_1 \neq x_2} E_\gamma(K(\cdot|x_1) \| K(\cdot|x_2)).\end{aligned}$$

Since  $E_\gamma(K(\cdot|x) \| K(\cdot|x)) = 0$  for every  $x \in \mathcal{D}$  and  $E_\gamma$  is non-negative, the desired inequality follows.

Let  $\gamma \in (0, 1)$ . Similar to (42), it can be shown that, for any probability measures  $\mu, \nu \in \mathcal{P}(\mathcal{D})$ ,

$$E_\gamma(\mu \| \nu) = \frac{1}{2} \|\mu - \gamma\nu\| - \frac{1}{2}(1 - \gamma).$$

A straightforward manipulation leads to

$$E_\gamma(\mu \| \nu) = \gamma \left( \frac{1}{2} \|\nu - (1/\gamma)\mu\| + \frac{1}{2}(1 - 1/\gamma) \right).$$

Since  $1/\gamma > 1$ , (42) implies that

$$E_\gamma(\mu \| \nu) = \gamma E_{1/\gamma}(\nu \| \mu). \quad (47)$$

In particular,  $E_\gamma(\mu \| \nu) = 0$  if and only if  $E_{1/\gamma}(\nu \| \mu) = 0$ . Therefore, (47) and the definition of  $\eta_\gamma(K)$  in (4) imply

$$\begin{aligned}\eta_\gamma(K) &= \sup_{\substack{\mu, \nu \in \mathcal{P}(\mathcal{D}): \\ E_\gamma(\mu \| \nu) \neq 0}} \frac{E_\gamma(\mu K \| \nu K)}{E_\gamma(\mu \| \nu)} \\ &= \sup_{\substack{\mu, \nu \in \mathcal{P}(\mathcal{D}): \\ E_{1/\gamma}(\nu \| \mu) \neq 0}} \frac{E_{1/\gamma}(\nu K \| \mu K)}{E_{1/\gamma}(\nu \| \mu)} \\ &= \eta_{1/\gamma}(K),\end{aligned}$$

as we wanted to show.

## B. Proof of Proposition 1

By Theorem 2, we have that

$$\begin{aligned}\eta_\gamma(K) &= \sup_{x_1, x_2 \in \mathcal{D}} E_\gamma(K(\cdot|x_1) \| K(\cdot|x_2)) \\ &= \sup_{x_1, x_2 \in \mathcal{D}} E_\gamma(\mathcal{N}(x_1, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(x_2, \sigma^2 \mathbf{I}_d)),\end{aligned}$$

It could be verified that  $r \mapsto \theta_\gamma(r)$  is increasing. Hence, by (15), we conclude that

$$\begin{aligned}\eta_\gamma(K) &= \sup_{x_1, x_2 \in \mathcal{D}} \theta_\gamma \left( \frac{\|x_2 - x_1\|}{\sigma} \right) \\ &= \theta_\gamma \left( \frac{\text{dia}(\mathcal{D})}{\sigma} \right),\end{aligned}$$

as desired.

### C. Proof of Lemma 4

Recall that  $X_1, \dots, X_n$  are i.i.d. random variables and  $\mathsf{K}_i$  is the conditional distribution of  $Z_i$  given  $(X_i, Z^{i-1})$ . Let  $P_{Z_1, \dots, Z_n}$  and  $Q_{Z_1, \dots, Z_n}$  be the distribution of  $Z^n$  when  $X^n \sim P_0^{\otimes n}$  and  $X^n \sim P_1^{\otimes n}$ , respectively. By Pinsker's inequality, we have that

$$\begin{aligned} \text{TV}^2(P_{Z^n}, Q_{Z^n}) &\leq \frac{1}{2} D_{\text{KL}}(P_{Z^n} \| Q_{Z^n}) \\ &= \frac{1}{2} \sum_{i=1}^n D_{\text{KL}}(P_{Z_i|Z^{i-1}} \| Q_{Z_i|Z^{i-1}} | P_{Z^{i-1}}), \end{aligned} \quad (48)$$

where the last equality follows from the chain rule for KL divergence. Observe that, for all  $z^i \in \mathcal{Z}^i$ ,

$$\begin{aligned} P_{Z_i|Z^{i-1}}(z_i|z^{i-1}) &= \int_{\mathcal{X}} \mathsf{K}_i(z_i|x_i, z^{i-1}) dP_0(x_i) \\ &= (P_0 \otimes \delta_{z^{i-1}})\mathsf{K}_i. \end{aligned}$$

*Mutatis mutandis*, we can show that

$$Q_{Z_i|Z^{i-1}}(z_i|z^{i-1}) = (P_1 \otimes \delta_{z^{i-1}})\mathsf{K}_i.$$

Since  $\mathsf{K}_i \in \mathcal{Q}_{\varepsilon, \delta}$ , (19) implies that, for all  $z^i \in \mathcal{Z}^i$ ,

$$\begin{aligned} D_{\text{KL}}((P_0 \otimes \delta_{z^{i-1}})\mathsf{K}_i \| (P_1 \otimes \delta_{z^{i-1}})\mathsf{K}_i) &\leq \varphi(\varepsilon, \delta) D_{\text{KL}}(P_0 \otimes \delta_{z^{i-1}} \| P_1 \otimes \delta_{z^{i-1}}) \\ &= \varphi(\varepsilon, \delta) D_{\text{KL}}(P_0 \| P_1). \end{aligned}$$

Therefore, (48) becomes

$$\text{TV}^2(P_{Z^n}, Q_{Z^n}) \leq \frac{n\varphi(\varepsilon, \delta)}{2} D_{\text{KL}}(P_0 \| P_1).$$

Plugging the previous inequality in (21), we obtain the desired result.

### D. Proof of Corollary 2

Fix  $\omega \in (0, 1]$  and consider two distributions  $P_0$  and  $P_1$  on  $\{-\omega^{-\frac{1}{k}}, 0, \omega^{-\frac{1}{k}}\}$  defined as

$$P_0(-\omega^{-\frac{1}{k}}) = \omega, \quad P_0(0) = 1 - \omega,$$

and

$$P_1(\omega^{-\frac{1}{k}}) = \omega, \quad P_1(0) = 1 - \omega.$$

It can be verified that both  $P_0$  and  $P_1$  belong to  $\mathcal{P}_k$ . Note that  $\ell_2^2(\theta(P_0), \theta(P_1)) = 2\omega^{\frac{2(k-1)}{k}}$ . Let  $M_0^n = P_0^{\otimes n} \mathsf{K}^n$  and  $M_1^n = P_1^{\otimes n} \mathsf{K}^n$  be the corresponding output distributions of the mechanism  $\mathsf{K}^n = \mathsf{K}_1 \dots \mathsf{K}_n$ , the composition of mechanisms  $\mathsf{K}_i$ . Le Cam's bound for  $\ell_2^2$ -metric yields

$$\begin{aligned} \mathcal{R}_n(\mathcal{P}_k, \ell_2^2, \varepsilon, \delta) &\geq \omega^{\frac{2(k-1)}{k}} (1 - \text{TV}(M_0^n, M_1^n)) \\ &\geq \omega^{\frac{2(k-1)}{k}} (1 - H(M_0^n, M_1^n)), \end{aligned} \quad (49)$$

where the last inequality follows from the fact  $\text{TV}(P, Q) \leq H(P, Q)$  for  $H(P, Q)$  being the Hellinger distance. Notice that  $M_0^n = \prod_{i=1}^n (P_0 \mathsf{K}_i)$  and  $M_1^n = \prod_{i=1}^n (P_1 \mathsf{K}_i)$  where each  $\mathsf{K}_i$  for  $i \in [n]$  is  $(\varepsilon, \delta)$ -LDP. It is well known that

$$H^2 \left( \prod_{i=1}^n P_i, \prod_{i=1}^n Q_i \right) = 2 - 2 \prod_{i=1}^n \left( 1 - \frac{1}{2} H^2(P_i, Q_i) \right).$$

Thus,

$$\begin{aligned}
H^2(M_0^n, M_1^n) &= 2 - 2 \prod_{i=1}^n \left( 1 - \frac{1}{2} H^2(P_0 K_i, P_1 K_i) \right) \\
&\leq 2 - 2 \prod_{i=1}^n \left( 1 - \frac{\varphi(\varepsilon, \delta)}{2} H^2(P_0, P_1) \right) \\
&= 2 - 2 \left( 1 - \frac{\varphi(\varepsilon, \delta)}{2} H^2(P_0, P_1) \right)^n \\
&= 2 - 2 (1 - \omega \varphi(\varepsilon, \delta))^n.
\end{aligned} \tag{50}$$

Hence, we obtain

$$\text{TV}(M_0^n, M_1^n) \leq \sqrt{2 - 2 (1 - \omega \varphi(\varepsilon, \delta))^n}. \tag{51}$$

Plugging (50) into (49), we obtain

$$\mathcal{R}_n(\mathcal{P}_k, \ell_2^2, \varepsilon, \delta) \geq \omega^{\frac{2(k-1)}{k}} \left[ 1 - \sqrt{2} \sqrt{1 - (1 - \omega \varphi(\varepsilon, \delta))^n} \right].$$

Now, choose  $\omega = \min \left\{ 1, \frac{1}{\varphi(\varepsilon, \delta)} \left[ 1 - \left( \frac{7}{8} \right)^{\frac{1}{\sqrt{n}}} \right] \right\}$ . Notice that we assume  $\delta > 0$  and hence  $\varphi(\varepsilon, \delta) > 0$  regardless of  $\varepsilon$ . Plugging this choice of  $\omega$  into the above bound, we obtain

$$\begin{aligned}
\mathcal{R}_n(\mathcal{P}_k, \ell_2^2, \varepsilon, \delta) &\gtrsim (\varphi(\varepsilon, \delta))^{-\frac{2(k-1)}{k}} \left[ 1 - \left( \frac{7}{8} \right)^{\frac{1}{\sqrt{n}}} \right]^{\frac{2(k-1)}{k}} \\
&\gtrsim (\varphi(\varepsilon, \delta))^{-\frac{2(k-1)}{k}} n^{-\frac{k-1}{k}}.
\end{aligned}$$

#### E. Proof of Lemma 5

Recall that  $X_1, \dots, X_n$  are i.i.d. random variables and  $K_i$  is the conditional distribution of  $Z_i$  given  $X_i$ . Note that the distribution of  $Z^n$  is  $P_0^{\otimes n} K^n$  or  $P_1^{\otimes n} K^n$  whenever  $X^n \sim P_0^{\otimes n}$  or  $X^n \sim P_1^{\otimes n}$ , respectively. By the definition of contraction coefficient, we have that

$$\begin{aligned}
\text{TV}(P_0^{\otimes n} K^n, P_1^{\otimes n} K^n) &\leq \eta_{\text{TV}}(K^n) \text{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \\
&\leq \eta_{\text{TV}}(K^n) \sqrt{\frac{n}{2} D_{\text{KL}}(P_0 \| P_1)},
\end{aligned}$$

where the last inequality is due to Pinsker's inequality and the tensorization of KL-divergence. Since in the non-interactive setting we have that  $K^n = K^{\otimes n}$ , Lemma 2 implies that

$$\begin{aligned}
\eta_{\text{TV}}(K^n) &\leq \max_{i \in [n]} \left[ 1 - \left( \frac{1 - \eta_{e^\varepsilon}(K_i)}{e^\varepsilon} \right)^n \right] \\
&\leq 1 - e^{-n\varepsilon} (1 - \delta)^n,
\end{aligned}$$

where the last inequality follows from Theorem 3 and the fact that  $K_i \in \mathcal{Q}_{\varepsilon, \delta}$  for every  $i \in [n]$ .

#### F. Proof of Theorem 4

Let  $\hat{\Theta} = \Psi(Z^n)$  be an estimate of  $\Theta$  for some  $\Psi$  and  $p_\zeta := P_{\Theta \hat{\Theta}}(\ell(\Theta, \hat{\Theta}) \leq \zeta)$  and  $q_\zeta := (P_\Theta P_{\hat{\Theta}})(\ell(\Theta, \hat{\Theta}) \leq \zeta)$ , i.e.,  $p_\zeta$  and  $q_\zeta$  correspond to the probability of the event  $\{\ell(\Theta, \hat{\Theta}) \leq \zeta\}$  under the joint and product distributions, respectively. By definition, we have for any  $\gamma \geq 1$

$$I_\gamma(\Theta; \hat{\Theta}) = \mathbb{E}_\gamma(P_{\Theta \hat{\Theta}} \| P_\Theta P_{\hat{\Theta}})$$

$$\begin{aligned}
&= \sup_{A \in \mathcal{T} \times \mathcal{T}} [P_{\Theta \hat{\Theta}}(A) - \gamma(P_{\Theta} P_{\hat{\Theta}})(A)] \\
&\geq p_{\zeta} - \gamma q_{\zeta} \\
&\geq p_{\zeta} - \gamma \mathcal{L}(\zeta),
\end{aligned}$$

where the last inequality follows from the inequality  $q_{\zeta} \leq \mathcal{L}(\zeta)$ . Indeed, observe that

$$\begin{aligned}
q_{\zeta} &= \int_{\mathcal{T}} \int_{\mathcal{T}} 1_{\{\ell(\theta, \hat{\theta}) \leq \zeta\}} P_{\Theta}(\mathrm{d}\theta) P_{\hat{\Theta}}(\mathrm{d}\hat{\theta}) \\
&\leq \sup_{t \in \mathcal{T}} \int_{\mathcal{T}} 1_{\{\ell(\theta, t) \leq \zeta\}} P_{\Theta}(\mathrm{d}\theta) \\
&= \mathcal{L}(\zeta).
\end{aligned}$$

Recalling that  $\Pr(\ell(\Theta, \hat{\Theta}) > \zeta) = 1 - p_{\zeta}$ , the above thus implies

$$\Pr(\ell(\Theta, \hat{\Theta}) > \zeta) \geq 1 - I_{\gamma}(\Theta; \hat{\Theta}) - \gamma \mathcal{L}(\zeta). \quad (52)$$

Since  $\mathbb{E}[\ell(\Theta, \hat{\Theta})] \geq \zeta \Pr(\ell(\Theta, \hat{\Theta}) \geq \zeta)$  by Markov's inequality, we can write by setting  $\gamma = e^{\varepsilon}$

$$\begin{aligned}
R_n^{\text{Bayes}}(P_{\Theta}, \ell, \varepsilon, \delta) &\geq \zeta \left[ 1 - I_{e^{\varepsilon}}(\Theta; \hat{\Theta}) - e^{\varepsilon} \mathcal{L}(\zeta) \right] \\
&\geq \zeta \left[ 1 - I_{e^{\varepsilon}}(\Theta; Z^n) - e^{\varepsilon} \mathcal{L}(\zeta) \right],
\end{aligned}$$

where the second inequality comes from the data processing inequality for  $I_{\gamma}$ . To further lower bound the right-hand side, we write

$$\begin{aligned}
I_{e^{\varepsilon}}(\Theta; Z^n) &= \int_{\mathcal{T}} \mathsf{E}_{e^{\varepsilon}}(P_{Z^n|\Theta=\theta} \| P_{Z^n}) P_{\Theta}(\mathrm{d}\theta) \\
&\leq \eta_{e^{\varepsilon}}(P_{Z^n|X^n}) \int_{\mathcal{T}} \mathsf{E}_{e^{\varepsilon}}(P_{X^n|\Theta=\theta} \| P_{X^n}) P_{\Theta}(\mathrm{d}\theta) \\
&= \eta_{e^{\varepsilon}}(P_{Z^n|X^n}) I_{e^{\varepsilon}}(\Theta; X^n),
\end{aligned}$$

where the inequality follows from the definition of contraction coefficient. When  $n = 1$ , we have  $\eta_{e^{\varepsilon}}(\mathsf{K}) \leq \delta$  as  $\mathsf{K} = P_{Z|X}$  is assumed to be  $(\varepsilon, \delta)$ -DP. For  $n > 1$ , we invoke (20) to obtain  $\eta_{e^{\varepsilon}}(P_{Z^n|X^n}) \leq \varphi_n(\varepsilon, \delta)$ .

#### G. Proof of Corollary 4

Let  $\beta_n(\alpha) := \beta_n^{\infty, 1}(\alpha)$  be the non-private trade-off between type I and type II error probabilities (i.e.,  $Z^n = X^n$ ). According to Chernoff-Stein lemma (see, e.g., [104, Theorem 11.8.3]), we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \beta_n(\alpha) = -D_{\text{KL}}(P_0 \| P_1). \quad (53)$$

Assume now that,  $Z^n$  is the output of  $\mathsf{K}^{\otimes n}$  for an  $(\varepsilon, \delta)$ -LDP mechanism  $\mathsf{K}$ . According to (53), we obtain that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \beta_n^{\varepsilon, \delta}(\alpha) = -\sup_{\mathsf{K} \in \mathcal{Q}_{\varepsilon, \delta}} D_{\text{KL}}(P_0 \mathsf{K} \| P_1 \mathsf{K}). \quad (54)$$

By (19), we obtain the desired result.

#### H. Proof of Theorem 5

Let  $\mu_1$  be the initial distribution of the iterative process and  $\{\mathbf{f}_t\} \stackrel{i}{\sim} \{\mathbf{f}'_t\}$  be a pair of neighboring collections of cost functions. In light of Definition 4 and inequality (31), we have that

$$\delta = \mathsf{E}_{e^{\varepsilon}}(\mu_i \mathsf{K}_{\mathbf{f}_i} \| \mu_i \mathsf{K}_{\mathbf{f}'_i}) \prod_{t=i+1}^n \eta_{e^{\varepsilon}}(\mathsf{K}_{\mathbf{f}_t}). \quad (55)$$

We begin by bounding  $\eta_\gamma(K_{f_t})$ . Each kernel  $K_{f_t}$  can be decomposed as

$$K_{f_t} = \Pi_{\mathcal{W}} \circ K_t \circ \Psi_{f_t},$$

where  $K_t$  is a  $\Psi_{f_t}(\mathcal{W})$ -constrained additive Gaussian kernel with noise magnitude given by  $\sigma_t^2$ , as defined in Section II. By Theorem 2, we have

$$\begin{aligned} \eta_{e^\varepsilon}(K_{f_t}) &= \sup_{w_1, w_2 \in \mathcal{W}} E_{e^\varepsilon}(K_t(\Psi_{f_t}(w_1)) \| K_t(\Psi_{f_t}(w_2))) \\ &= \sup_{w_1, w_2 \in \Psi_{f_t}(\mathcal{W})} E_{e^\varepsilon}(K_t(w_1) \| K_t(w_2)) \\ &= \eta_{e^\varepsilon}(K_t) \\ &= \theta_{e^\varepsilon} \left( \frac{\text{dia}(\Psi_{f_t}(\mathcal{W}))}{\sigma_t} \right), \end{aligned} \quad (56)$$

where the last equality comes from Proposition 1.

Next, we look at the first term in the RHS of (55). By Jensen's inequality, we can write

$$E_{e^\varepsilon}(\mu_i K_{f_i} \| \mu_i K_{f'_i}) \leq \int_{\mathcal{W}} E_{e^\varepsilon}(K_{f_i}(w) \| K_{f'_i}(w)) d\mu_i(w).$$

Observe that, for every  $w \in \mathcal{W}$ ,

$$K_{f_i}(w) \sim \Pi_{\mathcal{W}}(\Psi_{f_i}(w) + \sigma_i Z),$$

where  $Z \sim \mathcal{N}(0, \mathbf{I}_d)$ . A similar relation holds for  $K_{f'_i}$ . By the data processing inequality, we obtain that

$$E_{e^\varepsilon}(K_{f_i}(w) \| K_{f'_i}(w)) \leq E_{e^\varepsilon}(\Psi_{f_i}(w) + \sigma_i Z \| \Psi_{f'_i}(w) + \sigma_i Z).$$

Therefore, Lemma 3 implies that

$$E_{e^\varepsilon}(\mu_i K_{f_i} \| \mu_i K_{f'_i}) \leq \int_{\mathcal{W}} \theta_{e^\varepsilon} \left( \frac{\|\Psi_{f_i}(y) - \Psi_{f'_i}(y)\|}{\sigma_i} \right) d\mu_i(y).$$

Since  $r \mapsto \theta_\gamma(r)$  is increasing, we conclude that

$$E_{e^\varepsilon}(\mu_i K_{f_i} \| \mu_i K_{f'_i}) \leq \theta_{e^\varepsilon} \left( \frac{\psi}{\sigma_i} \right), \quad (57)$$

By plugging (56) and (57) into (55), we obtain the desired result.

### I. Proof of Theorem 6

Assume that  $\{f_t\}$  and  $\{f'_t\}$  are collections of cost functions such that  $\{f_t\} \stackrel{i}{\sim} \{f'_t\}$  for some  $i \in [n]$ . Let  $T$  be a uniform random variable over  $[n]$ . If  $\mu_{T+1}$  and  $\mu'_{T+1}$  are the distributions of the output of Algorithm 1 applied to  $\{f_t\}$  and  $\{f'_t\}$ , respectively, then

$$\begin{aligned} \mu_{T+1} &= \frac{1}{n} \sum_{t=1}^n \mu_1 K_{f_1} \dots K_{f_t}, \\ \mu'_{T+1} &= \frac{1}{n} \sum_{t=1}^n \mu_1 K_{f'_1} \dots K_{f'_t}. \end{aligned}$$

The convexity<sup>6</sup> of  $(\mu, \nu) \mapsto E_\gamma(\mu \| \nu)$  and Jensen's inequality imply that

$$E_{e^\varepsilon}(\mu_{T+1} \| \mu'_{T+1}) \leq \frac{1}{n} \sum_{t=1}^n E_{e^\varepsilon}(\mu_1 K_{f_1} \dots K_{f_t} \| \mu_1 K_{f'_1} \dots K_{f'_t}). \quad (58)$$

Recall that  $f_j = f'_j$  for all  $j \neq i$ . In particular, we have that  $\mu_1 K_{f_1} \dots K_{f_t} = \mu_1 K_{f'_1} \dots K_{f'_t}$  for all  $t < i$  and hence

$$E_{e^\varepsilon}(\mu_{T+1} \| \mu'_{T+1}) \leq \frac{1}{n} \sum_{t=i}^n E_{e^\varepsilon}(\mu_i K_{f_i} \dots K_{f_t} \| \mu_i K_{f'_i} \dots K_{f'_t}), \quad (59)$$

where  $\mu_i = \mu_1 K_{f_1} \dots K_{f_{i-1}}$ . As in the proof of Theorem 5, each summand can be bounded via inequality (31) to obtain

$$E_{e^\varepsilon}(\mu_{T+1} \| \mu'_{T+1}) \leq \frac{1}{n} \sum_{t=i}^n E_{e^\varepsilon}(\mu_i K_{f_i} \| \mu_i K_{f'_i}) \prod_{j=i+1}^t \eta_{e^\varepsilon}(K_{f_j}).$$

Furthermore, (56) and (57) imply that

$$E_{e^\varepsilon}(\mu_{T+1} \| \mu'_{T+1}) \leq \frac{1}{n} \sum_{t=i}^n \theta_{e^\varepsilon}\left(\frac{\psi}{\sigma_i}\right) \prod_{j=i+1}^t \theta_{e^\varepsilon}\left(\frac{\text{dia}(\Psi_{f_j}(\mathcal{W}))}{\sigma_j}\right),$$

from which, and Definition 4, we obtain (33). By further exploiting the monotonicity of  $r \mapsto \theta_\gamma(r)$ , we obtain that

$$\delta \leq \frac{1}{n} \sum_{t=i}^n \theta_{e^\varepsilon}\left(\frac{\psi}{\sigma}\right) \prod_{j=i+1}^t \theta_{e^\varepsilon}\left(\frac{D}{\sigma}\right).$$

A straightforward manipulation leads to

$$\begin{aligned} \delta &\leq \frac{1}{n} \theta_{e^\varepsilon}\left(\frac{\psi}{\sigma}\right) \sum_{t=0}^{n-i} \left[ \theta_{e^\varepsilon}\left(\frac{D}{\sigma}\right) \right]^t \\ &\leq \frac{1}{n} \theta_{e^\varepsilon}\left(\frac{\psi}{\sigma}\right) \left[ 1 - \theta_{e^\varepsilon}\left(\frac{D}{\sigma}\right) \right]^{-1}, \end{aligned}$$

as claimed in (34).

### J. Proof of Corollary 5

Define  $\Psi_x^{\text{SGD}}(w) = w - \eta \nabla \ell(w, x)$  for  $x \in \mathcal{X}$ . The PNSGD algorithm can then be expressed as

$$W_{t+1} = \Pi_{\mathcal{W}} \left( \Psi_{x_t}^{\text{SGD}}(W_t) + \eta \sigma Z_t \right).$$

Applying Theorem 6, we obtain that this algorithm is  $(\varepsilon, \delta)$ -DP for  $\varepsilon \geq 0$  and

$$\delta = \frac{1}{n} \theta_{e^\varepsilon}\left(\frac{\psi}{\eta \sigma}\right) \left[ 1 - \theta_{e^\varepsilon}\left(\frac{D}{\eta \sigma}\right) \right]^{-1}, \quad (60)$$

where

$$\psi = \sup_{x_1, x_2 \in \mathcal{X}} \sup_{w \in \mathcal{W}} \|\Psi_{x_1}^{\text{SGD}}(w) - \Psi_{x_2}^{\text{SGD}}(w)\|,$$

<sup>6</sup>For any convex function  $f$  on  $\mathbb{R}_+$ , its perspective, i.e.,  $(p, q) \mapsto qf\left(\frac{p}{q}\right)$ , is convex on  $\mathbb{R}_+^2$ . Since  $D_f(\mu \| \nu) = \mathbb{E}_\nu \left[ f\left(\frac{d\mu}{d\nu}\right) \right]$ , it follows that  $(\mu, \nu) \mapsto D_f(\mu \| \nu)$  is convex.

and  $D = \max_{t \in [n]} \text{dia}(\Psi_{x_t}^{\text{SGD}}(\mathcal{W}))$ . Since  $\ell(\cdot, x)$  is  $L$ -Lipschitz for all  $x \in \mathcal{X}$ , the triangle inequality implies

$$\begin{aligned}\|\Psi_{x_1}^{\text{SGD}}(w) - \Psi_{x_2}^{\text{SGD}}(w)\| &= \eta \|\nabla \ell(w, x_1) - \nabla \ell(w, x_2)\| \\ &\leq 2\eta L,\end{aligned}$$

and thus

$$\psi \leq 2\eta L. \quad (61)$$

We can also write for each  $t \in [n]$

$$\begin{aligned}\text{dia}(\Psi_{x_t}^{\text{SGD}}(\mathcal{W})) &= \sup_{w_1, w_2 \in \mathcal{W}} \|\Psi_{x_t}^{\text{SGD}}(w_1) - \Psi_{x_t}^{\text{SGD}}(w_2)\| \\ &= \sup_{w_1, w_2 \in \mathcal{W}} \|w_1 - \eta \nabla \ell(w_1, x_t) - w_2 + \eta \nabla \ell(w_2, x_t)\| \\ &\leq \|w_1 - w_2\| + \eta \|\nabla \ell(w_1, x_t) - \nabla \ell(w_2, x_t)\| \\ &\leq \text{dia}(\mathcal{W}) + 2\eta L,\end{aligned} \quad (62)$$

implying

$$D \leq \text{dia}(\mathcal{W}) + 2\eta L. \quad (63)$$

Plugging (61) and (63) into (60), we obtain the desired result.

### K. Missing proof of Section IV-B

We start showing that if  $f : \mathcal{W} \rightarrow \mathbb{R}$  is convex and  $\beta$ -smooth, then  $G(w) := w - \eta \nabla f(w)$  is 1-Lipschitz for  $\eta \leq \frac{2}{\beta}$ . To do so, observe that for any  $w_1, w_2 \in \mathcal{W}$ ,

$$\begin{aligned}\|G(w_1) - G(w_2)\|^2 &= \|w_1 - w_2 + \eta \nabla f(w_2) - \eta \nabla f(w_1)\|^2 \\ &= \|w_1 - w_2\|^2 + \eta^2 \|\nabla f(w_1) - \nabla f(w_2)\|^2 - 2\eta \langle \nabla f(w_1) - \nabla f(w_2), w_1 - w_2 \rangle \\ &\leq \|w_1 - w_2\|^2 + \eta \left( \eta - \frac{2}{\beta} \right) \|\nabla f(w_1) - \nabla f(w_2)\|^2 \\ &\leq \|w_1 - w_2\|^2,\end{aligned} \quad (64)$$

where the inequality in (64) follows from Lemma 3.11 in [122], which states that  $\langle \nabla f(w_1) - \nabla f(w_2), w_1 - w_2 \rangle \geq \frac{1}{\beta} \|\nabla f(w_1) - \nabla f(w_2)\|^2$ .

Replacing  $f(w)$  with  $\ell(w, x)$ , we obtain that the function  $w \mapsto \Psi_x^{\text{SGD}}(w)$  is 1-Lipschitz for all  $x \in \mathcal{X}$ . To obtain (36), we modify (62) in the proof of Corollary 5 as follows

$$\begin{aligned}D &= \max_{t \in [n]} \text{dia}(\Psi_{x_t}^{\text{SGD}}(\mathcal{W})) \\ &= \sup_{w_1, w_2 \in \mathcal{W}} \|\Psi_{x_t}^{\text{SGD}}(w_1) - \Psi_{x_t}^{\text{SGD}}(w_2)\| \\ &\leq \text{dia}(\mathcal{W}).\end{aligned} \quad (65)$$

Plugging this and (61) into (60), we obtain (36).

### L. Proof of Proposition 2

For ease of notation, let  $w^* \in \arg \min_{w \in \mathcal{W}} f(w)$  and

$$\Delta := \mathbb{E}[f(W_T)] - \inf_{w \in \mathcal{W}} f(w).$$

Since  $T$  is uniformly distributed on  $[n]$ , we have that

$$\mathbb{E}[\mathbf{f}(W_T) \mid W^n] = \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\mathbf{f}(W_t) \mid W^n],$$

where  $W^n = (W_1, \dots, W_n)$ . Therefore,

$$\Delta = \mathbb{E} \left[ \frac{1}{n} \sum_{t=1}^n \{\mathbf{f}(W_t) - \mathbf{f}(w^*)\} \right].$$

Since  $\mathbf{F}$  is a family of convex functions, the function  $\mathbf{f}(w) = \mathbb{E}[\mathbf{f}_t(w)]$  is convex. In particular, it follows that

$$\Delta \leq \mathbb{E} \left[ \frac{1}{n} \sum_{t=1}^n \langle \nabla \mathbf{f}(W_t), W_t - w^* \rangle \right]. \quad (67)$$

Define

$$G_t := \nabla \mathbf{f}_t(W_t) - \sigma_t Z_t.$$

Recall that  $\mathbf{f}_t$  and  $Z_t$  are independent of  $W_t$ . Hence, by differentiation under the integral sign,

$$\mathbb{E}[G_t \mid W_t] = \nabla \mathbf{f}(W_t).$$

As a consequence, we obtain that

$$\begin{aligned} \mathbb{E}[\langle G_t, W_t - w^* \rangle] &= \mathbb{E}[\langle \mathbb{E}[G_t \mid W_t], W_t - w^* \rangle] \\ &= \mathbb{E}[\langle \nabla \mathbf{f}(W_t), W_t - w^* \rangle]. \end{aligned}$$

Therefore, (67) becomes

$$\Delta \leq \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\langle G_t, W_t - w^* \rangle]. \quad (68)$$

Recall that, for every  $w \in \mathbb{R}^d$  and every  $w_0 \in \mathcal{W}$ ,

$$\|\Pi_{\mathcal{W}}(w) - w_0\| \leq \|w - w_0\|.$$

Thus, the update rule in (40) and the definition of  $G_t$  imply that

$$\begin{aligned} \|W_{t+1} - w^*\|^2 &\leq \|W_t - \eta_t G_t - w^*\|^2 \\ &= \|W_t - w^*\|^2 + \eta_t^2 \|G_t\|^2 - 2\eta_t \langle G_t, W_t - w^* \rangle. \end{aligned}$$

A direct manipulation shows that  $\mathbb{E}[\langle G_t, W_t - w^* \rangle]$  is bounded above by

$$\mathbb{E} \left[ \frac{\|W_t - w^*\|^2 - \|W_{t+1} - w^*\|^2}{2\eta_t} \right] + \frac{\eta_t(M^2 + \sigma_t^2 d)}{2},$$

where the last inequality uses the fact that

$$\mathbb{E}[\|G_t\|^2] \leq M^2 + \sigma_t^2 d.$$

Therefore, (68) becomes

$$\begin{aligned} \Delta &\leq \frac{1}{n} \sum_{t=1}^n \mathbb{E} \left[ \frac{\|W_t - w^*\|^2 - \|W_{t+1} - w^*\|^2}{2\eta_t} \right] + \frac{1}{n} \sum_{t=1}^n \frac{\eta_t(M^2 + \sigma_t^2 d)}{2} \\ &= \frac{1}{n} \sum_{t=1}^n \left\{ \frac{\mathbb{E}[\|W_t - w^*\|^2]}{2} \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \right\} + \frac{1}{n} \sum_{t=1}^n \frac{\eta_t(M^2 + \sigma_t^2 d)}{2}, \end{aligned}$$

where for ease of notation we define  $1/\eta_0 = 0$ . Since  $\eta_t = \frac{\text{dia}(\mathcal{W})}{M\sqrt{t}}$  and  $\|W_t - w^*\| \leq \text{dia}(\mathcal{W})$  for every  $t$ ,

$$\Delta \leq \frac{M\text{dia}(\mathcal{W})}{2\sqrt{n}} + \frac{1}{n} \sum_{t=1}^n \frac{\eta_t(M^2 + \sigma_t^2 d)}{2},$$

Finally, by the inequality  $\sum_{t=1}^n \frac{1}{\sqrt{t}} \leq 2\sqrt{n}$ , we conclude

$$\Delta \leq \frac{M\text{dia}(\mathcal{W})}{2\sqrt{n}} + \frac{M\text{dia}(\mathcal{W})}{\sqrt{n}} + \frac{d}{2n} \sum_{t=1}^n \eta_t \sigma_t^2,$$

as required.