Supplemental Material For "Differenctially Private Empirical Risk Minimization with Non-convex Loss Functions"

Di Wang [1] **Changyou Chen** [1] **Jinhui Xu** [1]

A. Background on Markov semigroups and Infinitestimal Generator

In order to be self-contained, in this section we introduce the background and some preliminaries of Markov diffusion process. We refer the reader to (Raginsky et al., 2017; Bakry et al., 2013; Chen et al., 2015) for more details.

Let $\{W_t\}_{t\geq 0}$ be a continuous-time homogeneous Markov process with values in $\mathbb{R}^d$, and $P = \{P_t\}_{t\geq 0}$ be the corresponding Markov semigroup. That is

$$P_s g(W_t) = \mathbb{E}[g(W_{s+t})|W_t]$$

for all $s, t \geq 0$ and all bounded measurable functions $g : \mathbb{R}^d \mapsto \mathbb{R}$. A Borel probability measure π is called stationary or invariant if $\int_{\mathbb{R}^d} P_t g d\pi = \int_{\mathbb{R}^d} g d\pi$ for all g and t. Each of P_t can be extended to a bounded linear operator on $L^2(\pi)$, such that $P_t g \geq 0$ whenever $g \geq 0$ and $P_t 1 = 1$ for all t. The infinitestimal generator of the semigroup is a linear operator $\mathcal{L}$ defined on a dense subspace $\mathcal{D}(\mathcal{L})$ of $L^2(\pi)$ such that for any $g \in \mathcal{D}(\mathcal{L})$, we have $\partial_t P_t g = \mathcal{L} P_t g$. Also, $\mathcal{L}$ can be defined as

$$\mathcal{L}g(W_t) := \lim_{h\to 0} \frac{P_h g(W_t) - g(W_t)}{h}.$$

The inifinitestimal generator $\mathcal{L}$ defines the Dirichlet form

$$\mathcal{E}(g) := -\int_{\mathcal{R}^d} g\mathcal{L}g d\pi.$$

Let P be a Markov semigroup with the unique invariant distribution π and the Dirichlet form $\mathcal{E}$. We say that π satisfies a Poincaré inequality with constant c if for all probability measures $\mu \ll \pi$, we have

$$\chi^2(\mu||\pi) \leq c\mathcal{E}(\sqrt{\frac{d\mu}{d\pi}}),$$

where $\chi^2(\mu||\pi) := \|\frac{d\mu}{d\pi} - 1\|^2_{L^2(\pi)}$ is the χ^2 divergence, and $\frac{1}{c} \leq \lambda$ with λ being the spectral gap

$$\lambda := \inf\{\frac{\mathcal{E}g}{\int_{\mathbb{R}^d} g^2 d\pi} : g \in C^2, g \neq 0, \int_{\mathbb{R}^d} g = 0\}.$$

We say that π satisfies a Logarithmic Sobolev inequality with constant c if, for all $\mu \ll \pi$,

$$D(\mu||\pi) \leq 2c\mathcal{E}(\sqrt{\frac{d\mu}{d\pi}}),$$

where $D(\mu||\pi) = \int d\mu \log \frac{d\mu}{d\pi}$ is the KL-divergence.

Consider a Markov process $\{W_t\}_{t\geq 0}$ with a unique invariant distribution π and the Dirichlet form $\mathcal{E}$ such that π satisfies a Logarithmic Sobolev inequality with constant c. Then, we have the following (Bakry et al., 2013):

*Equal contribution [1]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, USA. Correspondence to: Di Wang <dwang45@buffalo.edu>.

Proceedings of the 36th *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019.

1. Let $\mu_t := \mathcal{L}(W_t)$, then we have $D(\mu_t||\pi) \leq D(\mu_0||\pi)e^{-\frac{2t}{c}}$.
2. If $\mathcal{E}g = \alpha \int \|\nabla g\|^2 d\pi$ for some $\alpha > 0$, then, for any $\mu \ll \pi$, $\mathcal{W}_2(\mu, \pi) \leq \sqrt{2c\alpha D(\mu||\pi)}$.

Given a data set $D \in \mathcal{Z}^n$ with Langevin Monte Carlo Dynamic:

$$dW_t = -\nabla F(W_t, D)dt + \sqrt{2}dB_t. \tag{1}$$

If $\nabla F(\cdot, D)$ is Lipschitz, then the Gibbs measure $\pi_D(dw) \propto e^{-\beta F(w;D)}$ is the unique invariant measure of the underlying Markov semigroup. Its infinitestimal generator is

$$\mathcal{L}g(W_t) = (-\nabla F(W_t; D) \cdot \nabla + \Delta^2)g(W_t).$$

The corresponding Dirichlet from is

$$\mathcal{E}g = \int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi.$$

Under some assumptions about the loss function, (Raginsky et al., 2017) shows that the Gibbs measure satisfy logarithmic Sobolev inequality.

Lemma 1. *[Proposition 3.2 and Appendix B in (Raginsky et al., 2017)] For some $\beta \geq O(1)$, all of the Gibbs measures π satisfy a logarithmic Sobolev inequality with constant*

$$c_{LS} \leq O(\frac{1}{\lambda^*}(d + \beta)),$$

where λ_* is the uniform spectral gap

$$\lambda_* := \inf_{D \in \mathcal{Z}^n} \inf\{\frac{\int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi_D}{\int_{\mathbb{R}^d} g^2 d\pi_D} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_D), g \neq 0, \int_{\mathbb{R}^d} g d\pi_D = 0\}.$$

which satisfies:

$$\frac{1}{\lambda_*} \leq O(\frac{d + \beta}{\beta} \exp(O(\beta + d))).$$

Moreover, exponential dependence of $\frac{1}{\lambda}$ on β is unavoidable in the presence of multiple local minima and saddle points.

The following shows a connection between time average of the diffusion and the corresponding Poisson equation.

The Poisson equation is an elliptic PDE on the basis of the infinitestimal generator associated with the Langevin dynamics. For the generator $\mathcal{L}$ corresponding to the underlying Markov semigroup, we define the Poisson equation as

$$\mathcal{L}\psi = \phi - \bar{\phi},$$

where ϕ is the test function, and $\bar{\phi} := \int \phi(x)\pi(dx)$.

B. Omitted Proofs

B.1. Proof of Theorem 1

Proof. Firstly, we can see that if in each iteration

$$w_k = w_{k-1} + \eta(\nabla \hat{L}^r(w_{k-1}, D) + \xi_1) + \frac{\sqrt{\eta}}{\sqrt{\beta}}\xi_2,$$

where $\xi_1 \sim \mathcal{N}(0, \frac{L^2 c_2^2 \log(1/\delta)T}{n^2\epsilon^2} I_d)$ and $\xi_2 \sim \mathcal{N}(0, I_d)$, then by moment account (Lemma 1), we can see that it is (ϵ, δ)-differentially private for $\epsilon < c_1 T$ and $0 < \delta < 1$. Furthermore, if $\eta^2 \frac{L^2 c_2^2 \log(1/\delta)T}{n^2\epsilon^2} = \frac{\eta}{\beta}$ or $\eta = \frac{n^2\epsilon^2}{TL^2 c_2^2 \beta \log(1/\delta)}$, then it is equivalent to the updating in Algorithm 1. This completes the proof. □

B.2. Proof of Theorem 2

Proof of Theorem 2. The proof follows the framework of the proof in (Raginsky et al., 2017).

Notations For a given dataset D, we denote the corresponding Gibbs measure as $\pi_D \propto e^{-\beta \hat{L}^r(w,D)}$. Also, let $\mu_{T,D} = \mathcal{L}(w_T|D)$ and $\nu_{t,D} = \mathcal{L}(W_t|D)$.

Firstly, we show that our assumptions about the loss function and w_0 meet the assumptions in (Raginsky et al., 2017). Actually, our setting implies that $f(w,z) = \ell(w,z) + \frac{\lambda}{2}\|w\|^2$ in (Raginsky et al., 2017). It is easy to see that $A = A$, $B = L$, $M = M + \lambda$ in (Raginsky et al., 2017). Also, when $\ell(\cdot, z)$ is L-Lipschitz, we know that $f(w,z) = \ell(w,z) + \frac{\lambda}{2}\|w\|^2$ is $(m = \frac{\lambda}{2}, b = \frac{L^2}{2\lambda})$-dissipative (that is, $\langle w, \nabla f(w,z)\rangle \geq \frac{\lambda}{2}\|w\|^2 - \frac{L^2}{2\lambda}$) , which satisfies assumption A.3 in (Raginsky et al., 2017). For A.4, we can see that Algorithm 1 is just the non-stochastic version. Hence, $\delta = 0$. Thus, most of the analysis in (Raginsky et al., 2017) can also be applied here. For self-completeness, we will rephrase them so that they fit our differentially private context.

Now, we briefly introduce the proof in (Raginsky et al., 2017). Let $\hat{w}^*$ be the output of the Gibbs algorithm under which the conditional distribution of $\hat{w}^*$ is equal to π_D. Then, we decompose the population risk into the following

$$\mathbb{E}L^r_{\mathcal{P}}(w_T) - L^r_{\mathcal{P}}(w^*) = \mathbb{E}L^r_{\mathcal{P}}(w_T) - \mathbb{E}L^r_{\mathcal{P}}(\hat{w}^*) + \mathbb{E}L^r_{\mathcal{P}}(\hat{w}^*) - \mathbb{E}\hat{L}^r(\hat{w}^*, D) + \mathbb{E}\hat{L}^r(\hat{w}^*, D) - L^r_{\mathcal{P}}(w^*).$$

For the second term, by Proposition 3.5 in (Raginsky et al., 2017) we have

$$\mathbb{E}L^r_{\mathcal{P}}(\hat{w}^*) - \mathbb{E}\hat{L}^r(\hat{w}^*, D) \leq O(\frac{(\beta + d)c_{LS}}{n}) = O(\frac{\exp(O(\beta + d))}{n}) = O\big(\frac{\exp(O(\beta))}{n}\big), \tag{2}$$

where the big O notation hides the parameters of M, b, B (that is L, M, λ in our setting) by the assumption of $\beta > d$.

For the third term, we have the following theorem:

Lemma 2 ((Raginsky et al., 2017))**.** *For any $\beta \geq \frac{2}{m}$,*

$$\mathbb{E}\hat{L}^r(\hat{w}^*, D) - L^r_{\mathcal{P}}(w^*) \leq O(\frac{d}{\beta}\log(\beta)),$$

where the big O notation omits the factor of M, m.

In order to estimate the term of $\mathbb{E}L^r_{\mathcal{P}}(w_T) - \mathbb{E}L^r_{\mathcal{P}}(\hat{w}^*))$, we have to estimate $\mathbb{E}\hat{L}^r_D(w_T) - \mathbb{E}\hat{L}^r_D(\hat{w}^*))$ for each $D \in \mathcal{Z}^n$. The goal is to get an upper bound for $\mathcal{W}_2(\mu_{T,D}, \pi_D) \leq \mathcal{W}_2(\mu_T, \nu_{T\eta,D}) + \mathcal{W}_2(\nu_{T\eta,D}, \pi_D)$ for all dataset D.

For the term $\mathcal{W}_2(\nu_{T\eta,D}, \pi_D)$, since ν is related to the continuous-time Langevin diffusion (1), and $T\eta$ is a fixed value, which is independent of η, we have (see Section 3.4 in (Raginsky et al., 2017)):

$$\mathcal{W}_2(\nu_{T\eta,D}, \pi_D) \leq O\big(\sqrt{(d+\beta)c_{LS}}e^{-\frac{T\eta}{\beta c_{LS}}}\big) = O\big(\exp(O(\beta))\exp(-\frac{T\eta}{O(\exp(\beta))})\big). \tag{3}$$

Note that $T\eta = \frac{n^2\epsilon^2}{L^2 c_2^2 \beta \log(1/\delta)}$.

Our final goal is to estimate $\mathcal{W}_2(\mu_{T,D}, \nu_{T\eta,D})$. The proof is the same as in (Raginsky et al., 2017). However, we can see that $\frac{m}{4M^2} = \frac{\lambda}{8(M+\lambda)^2} \leq 1$. This means that in order to use the result in (Raginsky et al., 2017), we have to ensure that $\eta \leq O(\frac{m}{M^2}) = O(\frac{\lambda}{(M+\lambda)^2})$. That is, $T \geq C\frac{n^2\epsilon^2(M+\lambda)^2}{\beta L^2 \log(1/\delta)\lambda}$.

We can easily get (see Proposition 3.1 in (Raginsky et al., 2017)):

$$\mathcal{W}_2^2(\mu_{T,D}, \nu_{T\eta,D}) \leq O(\beta\sqrt{\eta}(T\eta)^2). \tag{4}$$

Thus, we have

$$\mathcal{W}_2(\mu_{T,D}, \pi_D) \leq O\big(\frac{(n\epsilon)^{\frac{5}{2}}}{\beta^{\frac{3}{4}}\log(1/\delta)T^{\frac{1}{4}}} + \sqrt{(d+\beta)c_{LS}}e^{-\frac{T\eta}{\beta c_{LS}}}\big). \tag{5}$$

For all $D \in \mathcal{Z}^n$, we have

$$\int \hat{L}^r(w, D)\mu_{T,D}(dw) - \int \hat{L}^r(w, D)\pi_D(dw) \leq O\Big(\frac{(n\epsilon)^{\frac{5}{2}}}{\beta^{\frac{3}{4}} \log(1/\delta) T^{\frac{1}{4}}} + \exp(O(\beta)) \exp(-\frac{n^2\epsilon^2}{\log(1/\delta) O(\exp(\beta))})\Big), \quad (6)$$

where O is independent of $\beta, T, n, \epsilon, \delta$.

Combining this with Lemmas 2, (6) and (2), we have the proof.

The result of the limit comes from the fact that $\exp(-x) \leq \frac{1}{x}$. □

B.3. Proof of Theorem 3

Proof of Theorem 3. For convenience, we let $F(w)$ denote $\hat{L}^r(w, D)$. Then, the updating becomes

$$w_{t+1} = w_t - \eta \nabla F(w_t) + \sqrt{\frac{2\eta}{\beta}} \zeta_t. \quad (7)$$

By scaling $\eta' = \frac{\eta}{\beta}$ and $F' = \beta F$, we have

$$w_{t+1} = w_t - \eta' \nabla F'(w_t) + \sqrt{2\eta'} \zeta_t. \quad (8)$$

Note that the technique of rescaling is commonly used in other papers, *e.g.,* (Dalalyan, 2017; Xu et al., 2018).

The continuous Langevin dynamic corresponding to (8) is

$$dW(t) = -\nabla F'(W(t))dt + \sqrt{2} dB(t). \quad (9)$$

$$\mathcal{L}g = -\nabla g \cdot \nabla F' + \Delta^2 g \quad (10)$$

Also the invariant distribution is $\pi(dw) \propto e^{-F'(W)}$, and the Poisson equation is

$$\mathcal{L}\psi = \phi - \bar{\phi}, \quad (11)$$

where $\bar{\phi} = \int \phi(w)\pi(dw)$ and ϕ is the testing function.

A seemingly straightforward way to prove the result is to use the theorem on finite time sample average error of SGLD, such as Theorem 2 in (Chen et al., 2015) or (55) in (Vollmer et al., 2016) to our equation (8). However, both papers consider only the case of $\beta = 1$, and their assumptions are quite strong compared to ours. This means that the hidden constants in their bounds may depend on β and the dimensionality d. However, as can be seen from above. β cannot be assumed as a constant in our problem. Thus we cannot directly apply their results.

Next, we will use some of the ideas in the proof of Theorem 9 in (Vollmer et al., 2016) to show that the the constants depend only polynomially on β and the degree of the polynomial is independent of d. We refer the reader to Section 9 in (Vollmer et al., 2016).

For convenience, we assume that the test function $\phi = F$. Now consider the solution ψ to the Possion equation (11) for ϕ (note that the existence will be shown later for a class ϕ of functions). Also, for $\psi(w_{t+1})$, we use Taylor expansion at w_t; that is (note that since we now only need to estimate the bias, we just expand it to the third order, which is different from the one in (Vollmer et al., 2016)),

$$\psi(w_{t+1}) = \psi(w_t) + \nabla\psi(w_t)(w_{t+1} - w_t) + \frac{1}{2}(w_{t+1} - w_t)^T \nabla^2 \psi(w_t)(w_{t+1} - w_t) + \mathcal{R}_t \quad (12)$$

$$= \psi(w_t) + \nabla\psi(w_t)(-\eta' \nabla F'(w_t) + \sqrt{2\eta'}\zeta_t) +$$
$$\frac{1}{2}\eta'^2 \nabla F'(w_t) \nabla'^2 \psi(w_t) \nabla F'(w_t) - \sqrt{2\eta'}\eta' \nabla F'(w_t)\zeta_t + \eta' \zeta_t^T \nabla^2 \psi(w_t)\zeta_t + \mathcal{R}_t, \quad (13)$$

where $\mathcal{R}_t = \frac{1}{6}\int_0^1 s^2 \psi^{(3)}(sw_t + (1-s)w_{t+1})(w_{t+1} - w_t, w_{t+1} - w_t, w_{t+1} - w_t)ds$ and (13) comes from (8).

Taking the expectation on $\psi(w_{t+1})$, we have

$$\mathbb{E}\psi(w_{t+1}) - \mathbb{E}\psi(w_t) = -\eta'\nabla\psi(w_t)\nabla F'(w_t) + \eta'\Delta^2\psi(w_t) + \frac{1}{2}\eta'^2\nabla F'(w_t)\nabla'^2\psi(w_t)\nabla F'(w_t) + \mathbb{E}\mathcal{R}_t. \tag{14}$$

By (10), we have

$$\eta'\mathbb{E}\mathcal{L}(\psi)(w_t) = \mathbb{E}\psi(w_{t+1}) - \mathbb{E}\psi(w_t) - \frac{1}{2}\eta'^2\nabla F'(w_t)\nabla'^2\psi(w_t)\nabla F'(w_t) - \mathbb{E}\mathcal{R}_t. \tag{15}$$

Summing over all t for $t = 1, \cdots, T$ and dividing $\eta' T$ on both sides, by Poisson equation (11) we get:

$$\mathbb{E}(\frac{\sum_{t=1}^T \phi(w_t)}{T} - \bar{\phi}) = \frac{1}{\eta' T}\mathbb{E}[\psi(w_{T+1}) - \psi(w_1)] - \frac{1}{\eta' T}\mathbb{E}\sum_{t=1}^T \mathcal{R}_t - \frac{1}{2}\frac{\eta'}{T}\sum_{t=1}^T \nabla F'(w_t)\nabla'^2\psi(w_t)\nabla F'(w_t). \tag{16}$$

What we need to prove are the following inequalities.

$$\sup_t \mathbb{E}\psi(w_t) \le C_1, \tag{17}$$

$$\sup_t \mathbb{E}\mathcal{R}_t \le \eta'^2 C_2 \tag{18}$$

$$\sup_t \mathbb{E}\nabla F'(w_t)\nabla'^2\psi(w_t)\nabla F'(w_t) \le C_3, \tag{19}$$

where C_1, C_2, C_3 are independent of η and at most polynomially depending on β with their degrees independent of d (note that they may depend on d, but we only care about β). If we can show these, then we have the proof.

To prove these inequalities, we want to show for the testing function ϕ and its corresponding ψ the following

$$\|\psi^{(i)}\| \le C_i^1 f, \forall i = \{0, 1, 2, 3\}, \tag{20}$$

where $\{C_i^1\}$ are constants that are at most polynomially depending on β (with degrees independent of d) and $f(x) = 1 + \|x\|_2^2$ is the quadratic function.

Also, we want to show for every $m \in \mathbb{N}$,

$$\sup_t \mathbb{E}\|w_t\|_2^m \le C_m^2 < \infty, \tag{21}$$

where $\{C_m^2\}$ are constants that also are at most polynomially depending on β (with degrees independent of d).

It is easy to see that if the above inequalities (*i.e.,* (20)(21)) can be proven, then for (17) we have $\sup_t \mathbb{E}\psi(w_t) \le O(C_0^1 C_2^2)$; for (18), we have

$$\begin{aligned}\sup_t \mathbb{E}\mathcal{R}_t \le& O\big(C_3^1(1 + \|w_t\|^2)[\eta'^3\|\nabla F'(w_t)\|^3 + \eta'^2\|\nabla F'(w_t)\|]\big)\\ \le& O(\eta'^2 C_3^1 f[\|\beta\nabla F(w_t)\|^3 + \|\beta\nabla F(w_t)\|]).\end{aligned} \tag{22}$$

Since F is smooth, we have $\|\nabla F(w)\| \le \frac{M_0}{2}(1 + \|w\|)$ for some M_0 independent of β. Thus, by (22), we have $\sup_t \mathbb{E}\mathcal{R}_t \le O(\eta'^2 C)$, where C is at most polynomially depending on β. Similarly, we can show for (19).

Thus, our goal is now to prove (20) and (21). For (21), we have the following theorem:

Theorem 1. *For every m, if $\beta > d$ and sufficiently small η in (7)*

$$\sup_t \mathbb{E}\|w_t\|_2^{2m} \le C_m^2 < \infty, \tag{23}$$

where w_t is in (7) and C_m^2 is independent of β.

Proof of Theorem 1. For m=1, it has been shown in Lemma 3.2 in (Raginsky et al., 2017) that $C_1^2 = O(\frac{d}{\beta}) = O(1)$, which satisfies our requirements. Actually, for any m, we can follow the proof of Lemma 3.2 in (Raginsky et al., 2017), to show that there is a sufficiently small η which makes the Theorem hold.

For example, when $m=2$, $\mathbb{E}\|w_t\|_2^4 = \mathbb{E}\|w_{t-1} - \eta\nabla F(w_{t-1}) + \sqrt{2\eta/\beta}\zeta_{t-1}\|^4 \leq O(\mathbb{E}\|w_{t-1} - \eta\nabla F(w_{t-1})\|^4 + \eta^4)$, also for $\mathbb{E}\|w_{t-1} - \eta\nabla F(w_{t-1})\|^4 \leq \mathbb{E}\|w_{t-1}\|^4(1 - \Theta(\eta) + \Theta(\eta^2) - \Theta(\eta^3) + \Theta(\eta^4)) + O(\eta)$. The constants in the big-Θ and big-O notations are independent of β. Thus, if we take a sufficiently small η, which makes $(1 - \Theta(\eta) + \Theta(\eta^2) - \Theta(\eta^3) + \Theta(\eta^4)) < 1$, then we can get an upper bound that is independent of β for $\beta \geq \max\{1, d\}$. The same argument goes for all $m \in \mathbb{N}$. Thus we have the proof. □

Proof of (21) Now by Theorem 1, we have for every m, $\sup_t \mathbb{E}\|w_t\|^m < C_m$, where C_m is at most polynomially depending on (actually is independent of) β, since by Jensen's Inequality we have $\sup_t \mathbb{E}\|w_t\|_2^m \leq \sqrt{\mathbb{E}\|w_t\|^{2m}}$. This proves (21).

Proof of (20) For (20), the key point is that our testing function is bounded by a quadratic function, due to the L-smoothness of our assumption. We have the following theorem due to Theorem 1 and 2 in (Pardoux & Veretennikov, 2001) (corresponding to the case of $\alpha = 1 > 0$, $b(x) = F'(x) = \beta L^r(w, D)$ and $r_0 = \infty$ in the Has'minski's assumption) and Theorem 13 in (Vollmer et al., 2016).

Theorem 2 (Theorem 1 and 2 in (Pardoux & Veretennikov, 2001))**.** *Consider the Poisson equation in $\mathbb{R}^d$,*

$$\mathcal{L}u(x) = -f(x), \tag{24}$$

where $\mathcal{L}$ is the infinitestimal generator of the diffusion process (9). We further assume that $\int f(x)\pi(dx) = 0$, where π is the invariant measure of the diffusion process. If $\|f(x)\| \leq C_1 + C_2\|x\|^s$ for some $s > 0$ and some constants C_1, C_2. Then (24) defines a continuous function $u(x)$ which belongs to the Sobolev class $W^2_{p,loc}$ for any $p > 1$, and satisfies the following properties,

1. *There exists a constant C' such that*
$$|u(x)| \leq C'(1 + \|x\|^s), \tag{25}$$
where C' is determined only by C_1, C_2 and C_m, and C_m is determined only by the constants in equations (4)-(6) in (Pardoux & Veretennikov, 2001) for $m > s + 2$.

2. *Moreover, there is a constant C such that*
$$\|\nabla u(x)\| \leq C(1 + \|x\|^s), \tag{26}$$
where C is determined only by C_1, C_2 and C_m, and C_m is determined only by the constants in equations (4)-(6) in (Pardoux & Veretennikov, 2001) for $m > s + 2$.

Now by the proofs of Theorem 1 and 2 in (Pardoux & Veretennikov, 2001), we have the following theorem:

Theorem 3. *For our test function ϕ, if fixing $m = 6$ in Theorem 2, then C' in (25) is polynomially depending on C_m, C_1, C_2, and the same for C in (26).*

Proof. The proof of C' depending polynomial on C_1, C_2, C_m can be easily found in the proof of Theorem 1 and Theorem 2 in (Pardoux & Veretennikov, 2001). Since for our test function ϕ, $s = 2$ by the M-smooth property. Thus, we need $m > \beta + 2$ and $m > 2k + 2$ for some $k > 0$ (See Proposition 1 in (Pardoux & Veretennikov, 2001)). This means that choosing $m = 6$ can satisfy the condition in Theorem 1 of (Pardoux & Veretennikov, 2001).

For C, we follows the proof of Theorem 1 in (Pardoux & Veretennikov, 2001). In (Pardoux & Veretennikov, 2001), the proof is by (25), Sobolev embedding theorem and Theorem 9.11 and (9.40) in (Gilbarg & Trudinger, 2015). From the proof of Theorem 9.11 in (Gilbarg & Trudinger, 2015), we can see that the hidden constant behind is only polynomially depending on the upper bounds of the coefficients of the second order PDE, which means only polynomially depending on β in our problem. Also by Sobolev embedding theorem, we can see that the polynomial dependence on β will be unchanged. Thus, we have the proof for C. □

Next, we show that C_1, C_2, C_m are at most polynomially depending on β.

Theorem 4. *For a fixed number m in (4)-(6) in (Pardoux & Veretennikov, 2001) related to the diffusion process (9), C_1, C_2 and the constants in (4)-(6) in (Pardoux & Veretennikov, 2001) are at most polynomially depending on β (which is r in assumption A_b in (Pardoux & Veretennikov, 2001)). Thus, C_m is at most polynomially depending on β.*

Proof. For C_1, C_2, since $f = \bar{\phi} - \phi$, which corresponds to (24) in Theorem 2, where $\phi = \hat{L}^r(\cdot, D)$, hence we have $\|f(x)\| \leq \|\hat{L}^r(x, D)\| + \|\bar{\phi}\|$. For the term of $\hat{L}^r(x, D)$, since it is $(M + \lambda)$-smooth, thus $\hat{L}^r(x, D) \leq \frac{M_0}{2}(1 + \|x\|^2)$ for some M_0 which is independent of β. For the term $\bar{\phi} = \int \hat{L}^r(w, D)\pi(dw)$, by Proposition 3.4 of (Raginsky et al., 2017), we know that if $\beta \geq \frac{2}{m} = \frac{4}{\lambda}$, then $\bar{\phi} \leq O(\frac{d}{\beta}\log(\beta + d) + \min \phi)$, which is at most polynomially depending on β. Thus, C_1, C_2 are at most polynomially depending on β with their degrees independent of the dimensionality d.

Now, let us consider C_m. Actually, by the proof of Theorem 1 in (Pardoux & Veretennikov, 2001), we can see that C_m only depends polynomially on the constants of (4)-(6) in proposition 1 in (Pardoux & Veretennikov, 2001). Thus, it suffices to show that constants of (4)-(6) in proposition 1 in (Pardoux & Veretennikov, 2001) depends only polynomially on β.

To show this, we can see that $\beta\frac{\lambda}{2}$ corresponds to r and $\alpha = 1$ in (Pardoux & Veretennikov, 2001). The proof of proposition 1 in (Pardoux & Veretennikov, 2001) comes from Lemma 1-Lemma 8 in (Veretennikov, 1997). From the proof in (Veretennikov, 1997), we know that all the constants of (4)-(6) in proposition 1 in (Pardoux & Veretennikov, 2001) are polynomially depending on r, *i.e.* β and their degrees are independent of d.

Thus C_1, C_2, C_m are all at most polynomially depending on β with their degrees independent of d.

□

To summarize, we have the following theorem:

Theorem 5. *The constant C and C' in (25) , (26) are at most polynomially depending on β, moreover, the degree of the polynomial is independent on d.*

Upto now, we have showed that $\|\psi^i\| \leq C_i^1 f$ for $i = \{0, 1\}$, where f is a quadratic function and C_i^1 are polynomially depending on β.

What is still left is for $i = \{2, 3\}$. To prove this, our idea is to use the trick in (Vollmer et al., 2016) (see A.9-A.11 and Lemma 15 in (Vollmer et al., 2016)). That is, we note that the derivatives of ψ can be expressed as the solution to different Poisson equations. Also, by iterating Theorem 2, 3, 4, 5, we can get all the constant C_i^1 depending at most polynomially on β.

Putting all these together, we have showed that

$$\mathbb{E}(\frac{\sum_{t=1}^T \phi(w_t)}{T} - \bar{\phi}) \leq C(\frac{1}{\eta' T} + \eta'), \tag{27}$$

where C is at most polynomially depending on β whose degree is independent of d (we omit other terms and consider only β). Taking $\eta' = \frac{\eta}{\beta}$ and η in Algorithm 1, also noting that $\mathbb{E}\hat{L}^r(w_j, D) = \mathbb{E}\frac{\sum_{i=1}^T \hat{L}^r(w_i, D)}{T}$ and $\phi = \hat{L}^r(\cdot, D)$, by Proposition 3.4 in (Raginsky et al., 2017), we can get the proof. □

B.4. Proof of Theorem 4

Lemma 3. *(Dwork et al., 2014) For the exponential mechanism $\mathcal{M}(D, u, \mathcal{R})$, we have*

$$Pr\{u(\mathcal{M}(D, u, \mathcal{R})) \leq OPT_u(x) - \frac{2\Delta u}{\epsilon}(\ln|\mathcal{R}| + t)\} \leq e^{-t}.$$

where $OPT_u(x)$ is the highest score in the range $\mathcal{R}$, i.e. $\max_{r\in\mathcal{R}} u(D, r)$.

We first show that the optimal value $w^* = \arg\min_{w\in\mathbb{R}^d} \hat{L}^r(w, D)$ contained in the ball $\mathbb{B}^d(\frac{L}{\lambda})$. This is because under our assumption, $\hat{L}^r(w, D)$ is $(\frac{\lambda}{2}, \frac{L^2}{2\lambda})$-dissipative. That is, $\forall w \in \mathbb{R}^d$, $\langle w, \nabla\hat{L}^r(w, D)\rangle \geq \frac{\lambda}{2}\|w\|^2 - \frac{L^2}{2\lambda}$. Thus, $w^* = \arg\min_{w\in\mathbb{B}^d(\frac{L}{\lambda})} \hat{L}^r(w, D)$.

For any $\alpha > 0$, by a simple volume argument (Lemma 5.2 in (Vershynin, 2010)) we can see that there exits an α-net $\mathcal{N}_\alpha$ whose size is at most $(1 + \frac{2L}{\lambda\alpha})^d$. Then, by the property that $\hat{L}^r(w, D)$ is $O(L)$-Lipschitz, we have the following:

$$\min_{w\in\mathcal{N}_\alpha} \hat{L}^r(w, D) - \hat{L}^r(w^*, D) \leq O(L\alpha).$$

Now consider the following ϵ-DP algorithm. We set the score function $u(D, w) = -(\hat{L}^r(w, D) - \hat{L}^r(w_0, D))$, where $w_0 \in \mathbb{B}^d(\frac{L}{\lambda})$ is an arbitrary point; the range space $\mathcal{R} = \mathcal{N}_\alpha$. Since $\hat{L}^r(w, D)$ is $O(L)$-Lipschitz in $\mathbb{B}^d(\frac{L}{\lambda})$, the sensitivity is at most $O(\frac{L}{n})$. Thus by Lemma 3 after running exponential mechanism, we have with probability at least $1 - \beta$,

$$\hat{L}^r(w^{\text{priv}}, D) - \min_{w \in \mathcal{N}_\alpha} \hat{L}^r(w, D) \leq O(\frac{d \ln \frac{1}{\alpha\beta}}{n\epsilon}).$$

Thus, form the above and taking $\alpha = \frac{d}{n\epsilon}$, we have

$$\hat{L}^r(w^{\text{priv}}, D) - \hat{L}^r(w^*, D) \leq \tilde{O}(\frac{d}{n\epsilon}).$$

Actually, this is the lower bound for ERM under general convex functions with the constrained set $\mathcal{C} = \mathcal{B}^d(r)$ under ϵ differential privacy, see Theorem 5.2 in (Bassily et al., 2014). By this, we can easily get a lower bound for non-convex loss functions under our assumptions. We thus have the following theorem:

Theorem 6. *Consider DP-ERM problem with $\hat{L}^r(w, D) = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i) + r(w)$, where $r(w) = \frac{\lambda}{2}\|w\|^2$, $\ell(w, z) = -\langle w, z\rangle - \frac{\lambda}{2}\|w\|^2$, $\mathcal{C} = \mathbb{B}^d(r)$ for some constant r. Then for every ϵ-differentially private algorithm, there is a dataset $D = \{z_1, \cdots, z_n\} \subseteq \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ such that, with probability at least 1/2, we must have:*

$$\hat{L}^r(w^{priv}, D) - \min_{w \in \mathcal{C}} \hat{L}^r(w, D) \geq \Omega(\min\{1, \frac{d}{n\epsilon}\}).$$

On the other hand, under our assumptions about $\mathcal{C} = \mathbb{B}^d(r)$, there is an ϵ-differentially private algorithm, whose output w^{priv} satisfies with probability at least $1 - \beta$,

$$\hat{L}^r(w^{priv}, D) - \hat{L}^r(w^*, D) \leq \tilde{O}(\frac{d}{n\epsilon}).$$

The time complexity is $O((1 + \frac{2Ln\epsilon}{d\lambda})^d n)$.

Thus we can get an near optimal bound for general non-convex loss functions under ϵ-differential privacy.

B.5. Proof of Theorem 5

Before showing the proof, we first give an upper bound on the Frank-Wolfe gap of the output in Algorithm 2 for general smooth and Lipschitz loss functions with general convex set $\mathcal{C}$. We start with the definition of Gaussian Width:

Definition 1 (Minkowski Norm)**.** *The Minkowski norm (denoted by $||\cdot||_C$) with respect to a centrally symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^d$ is defined as follows. For any vector $v \in \mathbb{R}^d$, $||\cdot||_C = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}$. The dual norm of $||\cdot||_C$ is denoted as $||\cdot||_{C^*}$; for any vector $v \in \mathbb{R}^d$, $||v||_{C^*} = \max_{w \in C} |\langle w, v\rangle|$.*

Definition 2 (Gaussian Width)**.** *Let $b \sim \mathcal{N}(0, I_d)$ be a Gaussian random vector in $\mathbb{R}^d$. The Gaussian width for a set $\mathcal{C}$ is defined as $G_C = \mathbb{E}_b[\sup_{w \in C}\langle b, w\rangle]$.*

Algorithm 1 DP-FW-L2

Input: T is the maximum of iterations, w_1 is the initial point, and $\{\gamma_t\}_{t=1}^T$ is the step size. ϵ and δ are privacy parameters.

for $t = 1, \cdots, T$ **do**
  Compute $v_t = \arg\max_{v \in \mathcal{C}}\langle v, -(\nabla L(w_t, D) + \epsilon_t)\rangle$, where $\epsilon_t \sim N(0, \sigma^2 I_d)$.
  $w_{t+1} = w_t + \gamma_t(v_t - w_t)$.
end for
Return $w_R \in \{w_1, \cdots, w_T\}$ such that R is uniformly sampled from $\{1, \cdots, T\}$.

Theorem 7. *Let C be a bounded, closed, centrally symmetric convex set. Assume that $\hat{L}(w, D)$ is differentiable and M-smooth over w with respect to ℓ_2 norm, and the loss function $\ell(\cdot, z)$ is L-Lipschitz over x with respect to ℓ_2-norm for all $z \in \mathcal{Z}$. Then, there are constants $c_1, c_2 > 0$ such that for any $0 < \epsilon < c_1 T, 0 < \delta < 1$, **DP-FW-L2** (Algorithm 2) is (ϵ, δ)-differentially private if $\sigma^2 = c_2 \frac{L^2 T \ln(\frac{1}{\delta})}{n^2 \epsilon^2}$. Moreover, if take $\{\gamma_t\}_{t=1}^T = O(\frac{\sqrt[4]{(\|C\|_2^2 + G_C^2) \ln \frac{1}{\delta}}}{\|C\|_2 \sqrt{n\epsilon}})$ and $T = O(\frac{n\epsilon}{\sqrt{(\|C\|_2^2 + G_C^2) \ln \frac{1}{\delta}}})$, the following holds,*

$$\mathbb{E}[\mathcal{G}_R] \leq O\Big(\frac{\|C\|_2 \sqrt[4]{(\|C\|_2^2 + G_C^2) \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\Big), \tag{28}$$

where $\mathcal{G}_t = \max_{v \in C} \langle -\nabla \hat{L}(w_t, D), v - w_t \rangle$.

Proof. To prove Theorem 7, we need the following lemmas.

Lemma 4. *For any vector v, we have $||v||_2 \leq ||C||_2 ||v||_C$, where $||C||_2$ is the ℓ_2-diameter and $||C||_2 = \sup_{x,y \in C} ||x - y||_2$.*

Lemma 4 implies that any smooth convex function $F(\theta)$, which is M-smooth with respect to ℓ_2 norm, is $M||C||_2^2$-smooth with respect to $||\cdot||_C$ norm, which is the motivation of our algorithm.

Proof. If $v = 0$, this is trivially true. Otherwise, we will show that $\frac{||v||_2}{||C||_2} \leq ||v||_C$. This is equivalent to show that $v \notin \frac{||v||_2}{||C||_2} C$. Taking any $y \in C$, since $||\frac{||v||_2}{||C||_2} y||_2 = \frac{||v||_2}{||C||_2} ||y||_2$, we know that $||y||_2 < ||C||_2$. Thus, $||\frac{||v||_2}{||C||_2} y||_2 < ||v||_2$. We have $v \notin \frac{||v||_2}{||C||_2} C$. □

Proof of Theorem 7. For convenience, we let the norm $\|\cdot\| = \|\cdot\|_C$, and $F(w) = \hat{L}(w, D)$. Let $\tilde{M}$ denote $M\|C\|_2^2$, and D denote the diameter of C w.r.t. $\|\cdot\|$ norm. By the M-smoothness property and Lemma 4, we have

$$F(w_{t+1}) \leq F(w_t) + \gamma_t \langle \nabla F(w_t), v_t - w_t \rangle + \frac{\tilde{M} \gamma_t^2}{2} \|v_t - w_t\|^2. \tag{29}$$

Let $\hat{v}_t = \arg\max_{v \in C} \langle v, -\nabla F(w_t) \rangle$. By the optimality of v_t, we have

$$\langle v_t, -\nabla F(w_t) - \epsilon_t \rangle \geq \langle \hat{v}_t, -\nabla F(w_t) - \epsilon_t \rangle.$$

This implies that

$$\langle v_t - \hat{v}_t, \nabla F(w_t) \rangle \leq \langle v_t - \hat{v}_t, -\epsilon_t \rangle. \tag{30}$$

From (29), we get

$$F(w_{t+1}) \leq F(w_t) + \gamma_t \langle \nabla F(w_t), v_t - \hat{v}_t \rangle + \gamma_t \langle \nabla F(w_t), \hat{v}_t - w_t \rangle + \frac{\gamma_t^2 \tilde{M}}{2} \|v_t - w_t\|^2.$$

Plugging (30) into (29) and by the fact that $\langle \nabla F(w_t), \hat{v}_t - w_t \rangle = -\mathcal{G}_t$ (from the definition of $\hat{v}_t$), we obtain

$$\begin{aligned}
\gamma_t \mathcal{G}_t &\leq F(w_t) - F(w_{t+1}) + \gamma_t \langle v_t - \hat{v}_t, -\epsilon_t \rangle + \frac{\tilde{M} \gamma_t^2}{2} D^2 \\
&\leq F(w_t) - F(w_{t+1}) + \frac{\gamma_t^2 \tilde{M} \|v_t - \hat{v}_t\|^2}{2} + \frac{\|\epsilon_t\|_*^2}{2\tilde{M}} + \frac{\tilde{M} \gamma_t^2}{2} D^2 \\
&\leq F(w_t) - F(w_{t+1}) + \frac{\|\epsilon_t\|_*^2}{2\tilde{M}} + \tilde{M} \gamma_t^2 D^2,
\end{aligned}$$

where the second inequality is due to Cauchy Inequality. By the definition of $\mathcal{G}_R$, we have $\mathbb{E}[\mathcal{G}_R] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathcal{G}_t]$. Since $\{\gamma_t\}_{t=1}^T = \gamma$, summing the above over $t = 1 \cdots, T$ and taking the expectation, we have

$$\mathbb{E}\mathcal{G}_R \leq \frac{F(w_1) - F(w^*)}{\gamma T} + \tilde{M} \gamma D^2 + \frac{1}{\gamma} O((\|C\|_2^2 + G_C^2) \frac{L^2 T \ln(\frac{1}{\delta})}{n^2 \epsilon^2}).$$

Taking $\gamma = O(\frac{\sqrt[4]{G^2(\|C\|_2^2+G_C^2)\ln\frac{1}{\delta}}}{\sqrt{\tilde{M}}D\sqrt{n\epsilon}})$ and $T = O(\frac{n\epsilon}{\sqrt{(\|C\|_2^2+G_C^2)L^2\ln\frac{1}{\delta}}})$, and by definition of $\|\cdot\|$ and the fact that $D \le O(1)$, we have the proof. □

We first consider the Generalized Linear Model. The following inequality has been proved in (Foster et al., 2018). We rephrase it here to make the proof self-complete. Denote by $L_{\mathcal{P}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{Z}}\ell(w;x,y)$ and $\hat{L}(w,D) = \frac{1}{n}\sum_{i=1}^n \ell(w;x,y)$.

Generalized Linear Model

Lemma 5. *For a fixed w, $L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \le \frac{C_\sigma}{2c_\sigma}\langle\nabla L_{\mathcal{P}}(w), w - w^*\rangle$.*

Proof. Let $w \in \mathcal{C}$ be fixed. Then, we have

$$\begin{aligned}\langle\nabla L_{\mathcal{P}}(w), w - w^*\rangle &= 2\mathbb{E}_{(x,y)}[\sigma(\langle w,x\rangle - y)\sigma'(\langle w,x\rangle)\langle w - w^*, x\rangle] \\ &= 2\mathbb{E}_x[(\sigma(\langle w,x\rangle) - \sigma(\langle w^*,x\rangle))\sigma'(\langle w,x\rangle)\langle w - w^*, x\rangle].\end{aligned}$$

By assumption 2, we have

$$\begin{aligned}L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) &= \mathbb{E}(\sigma(\langle w,x\rangle) - \sigma(\langle w^*,x\rangle))^2 \\ &\le \frac{C_\sigma}{2c_\sigma}\langle\nabla L_{\mathcal{P}}(w), w - w^*\rangle.\end{aligned}$$

□

By Lemma 5 and Theorem 7, we only need to bound $\langle\nabla L_{\mathcal{P}}(w), w - w^*\rangle$. Before doing that, we show that the empirical risk is Lipschitz and smooth, which satisfies the assumption in Theorem 7. It is due to:

$$\|\nabla\hat{L}(w,D)\|_2 = \|\frac{1}{n}\sum_{i=1}^n(\sigma(\langle w,x_i\rangle) - y_i)\sigma'(\langle w,x\rangle)x^T\|_2 \le C_\sigma(B+1), \tag{31}$$

and

$$\|\nabla^2\hat{L}(w,D)\|_2 = \|\frac{1}{n}\sum_{i=1}^n[(\sigma'(\langle w,x_i\rangle))^2 + \sigma''(\langle w,x_i\rangle)(\sigma(\langle w,x_i\rangle) - y_i)]x_ix_i^T\| \le C_\sigma^2 + C_\sigma(B+1).$$

Thus, $\hat{L}(w,D)$ is $(C_\sigma(B+1))$-Lipschitz and $C_\sigma^2+C_\sigma(B+1)$-smooth. Also, since $\mathcal{C}$ is the unit ℓ_2 norm, we have $G_C = O(\sqrt{d})$ and $\|C\|_2 = 1$. Thus, we get $\mathbb{E}[\mathcal{G}_R] \le O\big(\frac{\sqrt[4]{d\ln\frac{1}{\delta}}}{\sqrt{n\epsilon}}\big)$ by Theorem 7.

By the definition of $\mathcal{G}_R$, we know that $\mathbb{E}[\mathcal{G}_R] \ge \mathbb{E}\langle\nabla\hat{L}(w_R,D), w_R - w^*\rangle$. Taking the expectation w.r.t $\{(x_1,y_1),\cdots,(x_n,y_n)\}$, we then have $\mathbb{E}[\mathcal{G}_R] \ge \mathbb{E}\langle\nabla L_{\mathcal{P}}(w_R), w_R - w^*\rangle$. Combing it with Lemma 5, we get

$$\mathbb{E}L_{\mathcal{P}}(x_R) - L_{\mathcal{P}}(w^*) \le O\big(\frac{C_\sigma}{2c_\sigma}\frac{\sqrt[4]{d\ln\frac{1}{\delta}}}{\sqrt{n\epsilon}}\big).$$

Robust Regression We now consider robust regression. We begin with showing a similar result as in Lemma 5. First, the smoothness of ψ implies that for any $s, s^* \in \mathcal{S}$, we have

$$\psi(s) - \psi(s^*) \le \psi'(s^*)(s - s^*) + \frac{C_\psi}{2}(s - s^*)^2.$$

Taking $s = \langle w,x\rangle$ and $s^* = \langle w^*,x\rangle$, and then taking expectation w.r.t. $\{(x_1,y_1),\cdots,(x_n,y_n)\}$, we get

$$\begin{aligned}L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) &\le \mathbb{E}_{x,y}[\psi'(\langle w^*,x\rangle - y)\langle w - w^*, x\rangle] + \frac{C_\psi}{2}\mathbb{E}\langle w - w^*, x\rangle^2 \\ &= \langle\nabla L_{\mathcal{P}}(w^*), w - w^*\rangle + \frac{C_\psi}{2}\mathbb{E}\langle w - w^*, x\rangle^2.\end{aligned} \tag{32}$$

By Assumption 3, we have

$$\nabla L_{\mathcal{P}}(w^*) = \mathbb{E}_{x,\xi}[\psi'(-\xi)x] = 0.$$

Thus, we get $L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \leq \frac{C_\psi}{2}\mathbb{E}\langle w - w^*, x\rangle^2$. On the other hand, using gradient we have

$$\begin{aligned}\langle \nabla L_{\mathcal{P}}(w), w - w^*\rangle &= \mathbb{E}_x[\mathbb{E}_\xi \psi'(\langle w - w^*, x\rangle - \xi)\langle w - w^*, x\rangle] \\ &= \mathbb{E}_x[h(\langle w - w^*, x\rangle)\langle w - w^*, x\rangle].\end{aligned}$$

By the assumption on function $h(\cdot)$, we get

$$h(\langle w - w^*, x\rangle)\langle w - w^*, x\rangle = \frac{h(\langle w - w^*, x\rangle)}{\langle w - w^*, x\rangle}\langle w - w^*, x\rangle^2 \geq c_\psi \langle w - w^*, x\rangle^2,$$

where the inequality is due to the fact that $h(0) = 0$ and $h'(0) \geq c_\psi$.

Taking the expectation, we have

$$\langle \nabla L_{\mathcal{P}}(w), w - w^*\rangle \geq c_\psi \mathbb{E}_x\langle w - w^*, x\rangle^2.$$

Thus, we have the following lemma.

Lemma 6.

$$L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \leq \frac{C_\psi}{2c_\psi}\langle \nabla L_{\mathcal{P}}(w), w - w^*\rangle.$$

It is easily to get that the loss function $\ell(w, (x, y)) = \psi(\langle w, x\rangle - y)$ is C_ψ-Lipschitz and C_ψ-smooth. Using the same argument as in the proof for the case of Generalized Linear model, we get the proof. □

B.6. Proof of Theorem 6

We first give an upper bound on the Frank-Wolfe gap of general ℓ_1-norm Lipschitz and smooth loss functions.

Definition 3. *The loss function ℓ is L-Lipschitz under ℓ_1-norm over w, if for any $z \in \mathcal{Z}$ and $w_1, w_2 \in C$, $|\ell(w_1, z) - \ell(w_2, z)| \leq L||x_1 - x_2||_1$ holds.*

Definition 4. *A loss function $\ell : C \times \mathcal{Z} \mapsto \mathbb{R}$ is M-smooth over w with respect to the $||\cdot||_1$ norm if for any $z \in \mathcal{X}$ and $w_1, w_2 \in C$, the following holds*

$$||\nabla \ell(w_1, z) - \nabla \ell(x_2, z)||_\infty \leq M||w_1 - w_2||_1.$$

If f is differentiable, this yields $\ell(w_1, z) \leq \ell(w_2, z) + \langle \nabla \ell(w_2, z), w_1 - w_2\rangle + \frac{M}{2}||w_1 - w_2||_1^2$.

Assumption 1. *$\hat{L}(w, D)$ is assumed to be differentiable and M-smooth over x w.r.t ℓ_1-norm, and $\ell(\cdot, z)$ is assumed to be L-Lipschitz over x with respect to ℓ_1-norm for all $z \in \mathcal{X}$. $C \subseteq \mathbb{R}^d$ is assumed to be a closed convex set. Furthermore, C is assumed to be the convex hull of some finite set A, i.e., $C = Conv(A)$ and bounded. (For example, C could be a polytope.)*

Algorithm 2 DP-FW-L1

Input: T is the iteration number and x_1 is the initial point. $\{\gamma_t\}_{t=1}^T$ is the step size. $C \subseteq \mathbb{R}^p$ is the convex hull of a compact set $A \subseteq \mathbb{R}^d$. ϵ and δ are privacy parameters.

1: **for** $t = 1, \cdots, T$ **do**
2: Use exponential mechanism $\mathcal{M}(D, u, \mathcal{R})$, where $\mathcal{R} = A$, $u(D, s) = -\langle s, \nabla \hat{L}(w_t, D)\rangle$, to ensure $(\frac{\epsilon}{\sqrt{8T\ln(\frac{1}{\delta})}}, 0)$-differential privacy. Denote the output as $\tilde{w}_t$.
3: Compute $w_{t+1} = (1 - \gamma_t)w_t + \gamma_t \tilde{w}_t$.
4: **end for**
5: Return $w_R \in \{w_1, \cdots, w_T\}$, where R is uniformly sampled from $\{1, 2 \cdots, T\}$.

Theorem 8. *Under Assumption 1 and assuming that A is a finite set, then for any $\epsilon, \delta > 0$, **DP-FW-L1** (Algorithm 2) ensures (ϵ, δ)-differentially private. Furthermore, if set $T = O(\frac{n\epsilon}{\sqrt{\ln(\frac{1}{\delta})}\ln(|A|n/\eta)})$ and $\{\gamma_t\}_{t=1}^T = \sqrt{\frac{2}{MT\|C\|_1^2}}$, then with probability at least $1-\eta$, the following holds*

$$\mathbb{E}[\mathcal{G}_R] \leq O(\frac{\|C\|_1 \sqrt[4]{\ln(\frac{1}{\delta})}\sqrt{\ln\frac{n|A|}{\eta}}}{\sqrt{n\epsilon}}), \tag{33}$$

where $\mathcal{G}_t = \max_{v\in C}\langle -\nabla \hat{L}(w_t, D), v - w_t\rangle$.

Proof of Theorem 8. For convenience, we let $F(w) = \hat{L}(w, D)$. By exponential mechanism and advanced composition theorem, we can see that it is (ϵ, δ)-differentially private. By the L-Lipschitz (w.r.t ℓ_1-norm) property of the loss function, we know that $\Delta u \leq O(\frac{\|C\|_1 L}{n})$. Let $\beta = O(\frac{L\|C\|_1\sqrt{8T\ln(\frac{1}{\delta})}\ln(\frac{|A|T}{\eta})}{n\epsilon})$. By the utility bound of exponential mechanism (Lemma 3), we know that in each iteration, with probability $1 - \frac{\eta}{T}$, the following holds

$$\langle \tilde{w}_t, \nabla F(w_t)\rangle \leq \min_{v\in A}\langle v, \nabla F(w_t)\rangle + \beta. \tag{34}$$

Let $s_t = \arg\min_{u\in\mathcal{A}}\langle u, \nabla F(w_t)\rangle$. By the M-smooth property and (34), we have

$$\begin{aligned}
\frac{M}{2}\|w_{t+1} - w_t\|_1^2 &\geq F(w_{t+1}) - F(w_t) - \langle F(w_t), w_{t+1} - w_t\rangle \\
&= F(w_{t+1}) - F(w_t) - \gamma_t\langle \nabla F(w_t), \tilde{w}_t - w_t\rangle \\
&\geq F(w_{t+1}) - F(w_t) - \gamma_t(\langle \nabla F(w_t), s_t - w_t\rangle + \beta).
\end{aligned}$$

Note that $\min_{u\in C}\langle u - w_t, \nabla F(w_t)\rangle = \min_{u\in\mathcal{A}}\langle u - w_t, \nabla F(w_t)\rangle = \langle s_t - w_t, \nabla F(w_t)\rangle = -\mathcal{G}_t$. Thus, we have

$$F(w_{t+1}) - F(w_t) + \gamma_t\mathcal{G}_t \leq \gamma_t\beta + \frac{M\gamma_t^2}{2}\|C\|_1^2. \tag{35}$$

Summing over $t = 1, \cdots, T$, we get with probability $1-\eta$,

$$(\sum_{t=1}^T \gamma_t)\mathcal{G}_R \leq F(w_1) - F(w^*) + (\sum_{t=1}^T \gamma_t)\beta + \frac{M}{2}(\sum_{t=1}^T \gamma_t^2)\|C\|_1^2.$$

Taking $\{\gamma_t\}_{t=1}^T = \gamma$, we have

$$\mathcal{G}_R \leq \frac{F(w_1) - F(w^*)}{\gamma T} + \frac{\gamma\|C\|_1^2 M}{2} + O(\frac{L\|C\|_1\sqrt{T\ln(\frac{1}{\delta})}\ln(\frac{|A|T}{\eta})}{n\epsilon}).$$

Taking $T = O(\frac{n\epsilon}{L\sqrt{\ln(\frac{1}{\delta})}\ln(|A|n)})$ and $\gamma = \sqrt{\frac{2}{T\|C\|_1^2 M}}$, we get the result. □

Generalized Linear Model We first show the Lipschitz and Smooth properties w.r.t ℓ_1-norm. Since $\nabla\ell(w, x, y) = \sigma(\langle w, x\rangle - y)\sigma'(\langle w, x\rangle)x^T$, by the Lipschitzness and the assumption, we have

$$\|(\sigma(\langle w, x\rangle) - y)\sigma'(\langle w, x\rangle)x^T\|_\infty \leq C_\sigma(B+1).$$

Let $w_1, w_2 \in C$, we have

$$\begin{aligned}
&\|(\sigma(\langle w_1, x\rangle) - y)\sigma'(\langle w_1, x\rangle)x^T - (\sigma(\langle w_2, x\rangle) - y)\sigma'(\langle w_2, x\rangle)x^T\|_\infty \\
&\leq |(\sigma(\langle w_1, x\rangle) - y)\sigma'(\langle w_1, x\rangle) - (\sigma(\langle w_2, x\rangle) - y)\sigma'(\langle w_2, x\rangle)| \\
&\leq |\sigma(\langle w_1, x\rangle)\sigma'(\langle w_1, x\rangle) - \sigma(\langle w_2, x\rangle)\sigma'(\langle w_2, x\rangle)| + |\sigma'(\langle w_1, x\rangle) - \sigma'(\langle w_2, x\rangle)| \\
&\leq (C_\sigma^2 + (B+1)C_\sigma)|\langle w_1 - w_2, x\rangle| \\
&\leq (C_\sigma^2 + (B+1)C_\sigma)\|w_1 - w_2\|_1.
\end{aligned}$$

Thus, by Theorem 8, we know $\mathbb{E}[\mathcal{G}_R] \leq O(\frac{\sqrt[4]{\ln(\frac{1}{\delta})}\sqrt{\ln\frac{np}{\eta}}}{\sqrt{n\epsilon}})$. The remaining part of the proof is by Lemma 5 and is the same as in the proof of Theorem 5.

Robust Regression For the case of linear regression, it is almost the same as in the case of generalized linear model, we omit it here.

B.7. Proof of Theorem 7

The guarantee of (ϵ, δ)- DP comes from Lemma 1. Below we show that one of $\{w_1, w_2, \cdots, w_T\}$ is α-SOSP with high probability.

For convenience, we use the following notations $F(w) = \hat{L}(w, D)$, $\{\eta_t\} = \eta = \frac{1}{M}$, $\Phi = \sqrt{\frac{\alpha^3}{\rho}}\chi^{-3}c^{-5}$, $r = \alpha\chi^{-3}c^{-6}$, $\Gamma = \frac{\chi c}{\eta\sqrt{\rho\alpha}}$ and $\chi = \max\{1, C_1 \log\frac{dMB}{\rho\alpha\xi}\}$ for some constant C_1 and enough large constant c.

By the concentration inequality of Gaussian distribution, we have the following lemma.

Lemma 7. *With probability at least* $1 - \frac{\xi}{2}$*, for all* $i \in [T]$*,*

$$\|\epsilon_i\|_2 \leq \frac{\sqrt{2c_2 \log\frac{1}{\delta} T d L \log\frac{4T}{\xi}}}{n\epsilon} \leq r = \alpha\chi^{-3}c^{-6}.$$

Below we assume that the event in Lemma 7 happens. Next, we show the following.

Lemma 8. *If* $\|F(w_t)\|_2 \geq \alpha$*, then we have*

$$F(w_{t+1}) - F(w_t) \leq -\eta\frac{\alpha^2}{4}.$$

Proof of Lemma 8. By the M-smoothness and taking $\eta = \frac{1}{M}$, we have

$$\begin{aligned}
F(w_{t+1}) &\leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t\rangle + \frac{M}{2}\|w_{t+1} - w_t\|_2^2 \\
&\leq F(w_t) - \eta\|\nabla F(w_t)\|_2^2 + \eta\|\nabla F(w_t)\|_2\|\epsilon_t\|_2 + \frac{\eta^2 M}{2}[\|\nabla F(w_t)\|_2^2 + 2\|\nabla F(w_t)\|_2\|\epsilon_t\|_2 + \|\epsilon_t\|_2^2] \\
&= F(w_t) - \eta\|\nabla F(w_t)\|_2[\frac{1}{2}\|\nabla F(w_t)\|_2 - 2\|\epsilon_t\|_2] + \frac{\eta}{2}\|\epsilon_t\|_2^2 \\
&\leq F(w_t) - \frac{\eta\alpha^2}{4},
\end{aligned}$$

where the last inequality is due to the following: by the assumption on n, we have $\|\epsilon_t\| \leq \alpha\xi^{-3}c^{-6} \leq \frac{\alpha}{20}$ for sufficiently large c and $\|\nabla F(w_t)\|_2 \geq \alpha$. □

Next, we prove the following key lemma:

Lemma 9. *If* $\|\nabla F(w_t)\| \leq \alpha$ *and* $\lambda_{\min}(\nabla^2 F(w_t)) \leq -\sqrt{\rho\alpha}$*, then in Algorithm 4, with probability* $1 - \xi$*, we have* $F(w_{t+\Gamma}) - F(w_t) \leq -\Phi$.

Proof of Lemma 9. To prove this lemma, we need the following lemmas.

Lemma 10.

$$F(w_{t+1}) - F(w_t) \leq -\frac{\eta}{4}\|\nabla F(w_t)\|_2^2 + 5\eta\|\epsilon_t\|_2^2.$$

Proof of Lemma 10. By the M-smoothness, we have

$$\begin{aligned}
F(w_{t+1}) &\le F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{M}{2}\|w_{t+1} - w_t\|_2^2 \\
&\le F(w_t) - \eta \langle F(w_t), F(w_t) + \epsilon_t \rangle + \frac{\eta^2 M}{2}(\|\nabla F(w_t)\|_2^2 + 2\|\nabla F(w_t)\|_2\|\epsilon_t\|_2 + \|\epsilon_t\|_2^2) \\
&\le F(w_t) - \frac{\eta}{2}\|\nabla F(w_t)\|_2^2 + 2\eta\|\nabla F(w_t)\|_2\|\epsilon_t\|_2 + \frac{\eta}{2}\|\epsilon_t\|_2^2 \\
&\le F(w_t) - \frac{\eta}{4}\|\nabla F(w_t)\|_2^2 + 5\eta\|\epsilon_t\|_2^2.
\end{aligned}$$

□

Lemma 11. *For all $t + 1 \le T$, we have*

$$\|w_{t+1} - w_1\|_2^2 \le 8\eta T(F(w_1) - F(x_{T+1})) + 50\eta^2 T \sum_{t=1}^{T} \|\epsilon_t\|_2^2.$$

Proof of Lemma 11. For any $t \le T - 1$, by Lemma 10, we have

$$\begin{aligned}
\|w_{t+1} - w_t\|_2^2 &\le \eta^2\|\nabla F(w_t) + \epsilon\|_2^2 \le 2\eta^2\|\nabla F(w_t)\|_2^2 + 2\eta^2\|\epsilon_t\|_2^2 \\
&\le 8\eta(F(w_t) - F(w_{t+1})) + 50\eta^2\|\epsilon_t\|_2^2.
\end{aligned}$$

Thus we have

$$\sum_{t=1}^{T} \|w_{t+1} - w_t\|_2^2 \le 8\eta(F(w_1) - F(w_{T+1})) + 50\eta^2 \sum_{t=1}^{T} \|\epsilon_t\|_2^2.$$

In total, we get

$$\|w_{t+1} - w_1\|_2^2 \le (\sum_{i=1}^{t} \|w_{i+1} - w_i\|_2)^2 \le t \sum_{i=1}^{t} \|w_{i+1} - w_i\|_2^2 \le T \sum_{t=1}^{T} \|w_{t+1} - w_t\|_2^2.$$

□

Let $\mathrm{DPGD}_{\epsilon}^{(t)}(x)$ be our algorithm that updates t-times with perturbations $\{\epsilon_1, \cdots, \}$ fixed and begins with x. Define the stuck region as:

$$\mathcal{X}^{\epsilon}(\tilde{w}) = \{w | w \in \mathbb{B}_{\tilde{w}}(\eta r), \text{ and } Pr(F(DPGD_{\epsilon}^{(\Gamma)}(w)) - F(\tilde{w}) \ge -\Phi) \ge \sqrt{\xi}\}. \tag{36}$$

Intuitively, the later perturbations of the coupling sequence are the same while the very first perturbation is used to escape the saddle points.

Lemma 12. *There exists a large enough constant c such that if $\|\nabla F(\tilde{w})\|_2 \le \alpha$ and $\lambda_{\min}(\nabla^2 F(\tilde{w})) \le -\sqrt{\rho\epsilon}$, then the width of $\mathcal{X}^{\epsilon}(\tilde{w})$ along the minimum eigenvector of $\tilde{w}$ is at most $\xi\eta r\sqrt{\frac{2\pi}{d}}$.*

Proof of Lemma 12. To prove this lemma, we let $\boldsymbol{e}_{\min}$ be the minimum eigenvector of $\nabla^2 F(\tilde{w})$. It suffice to show that for any $w_1, w_1' \in \mathbb{B}_{\tilde{w}}(\eta r)$ satisfying the condition of $w_1 - w_1' = \lambda \boldsymbol{e}_{\min}$, where $|\lambda| \ge \xi\eta r\sqrt{\frac{2\pi}{d}}$, $w_1 \notin \mathcal{X}^{\epsilon}(\tilde{w})$ or $w_1' \notin \mathcal{X}^{\epsilon}(\tilde{w})$.

Let $w_{\Gamma+1} = DPGD_{\epsilon}^{(\Gamma)}(w_1)$ and $w_{\Gamma+1}' = DPGD_{\epsilon}^{(\Gamma)}(w_1')$, where the two sequences are independent. To show that $w_1 \notin \mathcal{X}^{\epsilon}(\tilde{w})$ or $w_1' \notin \mathcal{X}^{\epsilon}(\tilde{w})$, it is sufficient to demonstrate that with probability at least $1 - \xi$

$$\min\{F(w_{\Gamma+1}) - F(\tilde{w}), F(w_{\Gamma+1}') - F(\tilde{w})\} \le -\Phi. \tag{37}$$

That is due to the fact that if $w_1, w_1' \in \mathcal{X}^{\epsilon}(\tilde{w})$, we have, with probability at least ξ, that $F(w_{\Gamma+1}) - F(\tilde{w}) \ge -\Phi$ and $F(w_{\Gamma+1}') - F(\tilde{w}) \ge -\Phi$. This will mean that with probability at most $1 - \xi$,

$$\min\{F(w_{\Gamma+1}) - F(\tilde{w}), F(w_{\Gamma+1}') - F(\tilde{w})\} \le -\Phi, \tag{38}$$

which contradicts (37).

To prove that (37) holds with probability at least $1-\xi$, we need to show that

1. $\max\{F(w_1)-F(\tilde{w}), F(w_1')-F(\tilde{w})\} \le \Phi$,
2. with probability at least $1-\delta$, $\min\{F(w_{\Gamma+1})-F(w_1), F(w_{\Gamma+1}')-F(w_1')\} \le -2\Phi$.

For (1), we have, by the definition of $w_1 \in \mathbb{B}_{\tilde{w}}(\eta t)$ and the M-smoothness, that

$$F(w_1)-F(\tilde{w}) \le \alpha\eta r + \frac{M}{2}(\eta r)^2 = O(\frac{\alpha^2}{M}\chi^{-3}c^{-6}) \le \Phi$$

for sufficiently large c. Similarly, we have the same for $F(w_1')-F(\tilde{w})$.

To prove (2), we first assume that it is not true, *i.e.,*

$$\min\{F(w_{\Gamma+1})-F(w_1), F(w_{\Gamma+1}')-F(w_1')\} \ge -2\Phi.$$

Then, by Lemmas 7 and 11, we have $\forall t \in [\Gamma+1]$ that for sufficiently large $c>0$,

$$\begin{aligned}\max\{\|w_t-\tilde{w}\|_2, \|w_t'-\tilde{w}\|_2\} &\le \max\{\|w_t-w_1\|_2 + \|w_1-\tilde{w}\|_2, \max\{\|w_t'-w_1'\|_2+\|w_1'-\tilde{w}\|_2\}\}\\ &\le \sqrt{16\eta\Gamma\Phi + 50\eta^2\Gamma^2 r^2} + \eta r\\ &\le \sqrt{16\eta\Gamma\Phi + 50\eta^2\Gamma^2\alpha^2\chi^{-4}c^{-12}} + \eta\alpha\chi^{-3}c^{-6}\\ &\le 4(\sqrt{\frac{\alpha}{\rho}}\chi^{-1}c^{-2}) = R,\end{aligned}$$

where the last inequality is due to the fact that $M \ge \sqrt{\rho\alpha}$. This means that both sequences $\{w_t\}_{t=1}^{\Gamma+1}$ and $\{w_t'\}_{t=1}^{\Gamma+1}$ do not leave the ball with radius R around $\tilde{w}$. Let $H = \nabla^2 F(\tilde{w})$ and $x_t := w_t - w_t'$. We have

$$\begin{aligned}x_{t+1} &= x_t - \eta[\nabla F(w_t) - \nabla F(w_t')] = (I-\eta H)x_t - \eta\Delta_t x_t\\ &\le (I-\eta H)^t x_1 - \eta\sum_{\tau=1}^{t}(I-\eta H)^{t-\tau}(\Delta_\tau x_\tau)),\end{aligned}$$

where $\Delta_t = \int_0^1 [\nabla F(w_t' + \theta(w_t-w_t')) - H]d\theta$. By Hessian Lipshitz, we have $\Delta_t \le \rho\max\{\|w_t-\tilde{w}\|_2, \|w_t'-\tilde{w}\|_2\} \le \rho R$. We now show the following by induction:

$$\|\eta\sum_{\tau=1}^{t}(I-\eta H)^{t-\tau}(\Delta_\tau x_\tau)\| \le \frac{1}{2}\|(I-\eta H)^t x_1\|_2. \tag{39}$$

For the base case of $t=1$, we can easily verify it using the fact that $\eta\rho R \le \frac{1}{2}$ for sufficiently large c. Suppose that it holds for all $t' \le t$. This gives us $\|x_{t'}\|_2 \le 2\|(I-\eta H)^{t'}x_1\|_2$. Let $\gamma = \lambda_{\min}(\nabla^2 F(\tilde{w}))$. For the case of $t+1 \le \Gamma+1$, we have

$$\|\eta\sum_{\tau=1}^{t}(I-\eta H)^{t-\tau}(\Delta_\tau x_\tau)\| \le \eta\rho R\|\sum_{\tau=1}^{t}(I-\eta H)^{t-\tau}\|_2\|x_\tau\|_2 \le \eta\rho R\Gamma(1+\eta\gamma)^t\|x_1\|_2 \le \frac{1}{4}\|(I-\eta H)^t x_1\|_2,$$

where the third inequality uses the fact that x_0 is along the direction of the minimum eigenvector of H, and the last one is due to the fact that $\eta\rho R\Gamma = 4c^{-1} \le \frac{1}{4}$ for large enough constant c.

Thus, in total we have

$$\begin{aligned}
\|x_{\Gamma+1}\| &\geq \|(I-\eta H)^{\Gamma}x_1\|_2 - \|\eta\sum_{\tau=1}^{\Gamma}(I-\eta H)^{t-\tau}(\Delta_\tau x_\tau)\|_2 \\
&\geq \frac{1}{2}\|(I-\eta H)^{\Gamma}x_1\|_2 = \frac{1}{2}(1-\eta\gamma)^{\Gamma}\|x_1\|_2 \\
&\geq \frac{1}{2}(1+\eta\sqrt{\rho\epsilon})^{\Gamma}\|x_1\|_2 \\
&\geq \frac{1}{2}(1+\eta\sqrt{\rho\epsilon})^{\Gamma}\xi\eta r\sqrt{\frac{2\pi}{d}} \\
&\geq (1+\eta\sqrt{\rho\alpha})^{\Gamma}\frac{\xi\alpha\chi^{-3}c^{-6}}{2M} \\
&\geq 2^{\eta\sqrt{\rho\alpha}\Gamma}\frac{\xi\alpha\chi^{-3}c^{-6}}{2M} \\
&\geq 8\sqrt{\frac{\alpha}{\rho}}\chi^{-1}c^{-2} = 2R,
\end{aligned}$$

where the last inequality is due to the fact that $\Gamma = \frac{\chi c}{\eta\sqrt{\rho\alpha}}$ and $\chi = \max\{1, \log\frac{\sqrt{d}M}{\xi\sqrt{\rho\alpha}}\}$. From the above, we can see that when c is sufficiently large, the above inequalities hold. Thus, we have $\|x_{\Gamma+1}\|_2 \geq 2R$. This contradicts the fact that $\max\{\|w_t - \tilde{w}\|_2, \|w_t' - \tilde{w}\|_2\} \leq R$. This completes the proof. □

We now return to the proof of Lemma 9. Let $r_0 = \xi r\sqrt{\frac{2\pi}{d}}$. By Lemma 12, we know that $\mathcal{X}^{\epsilon}(w_t)$ has width at most ηr_0 in the direction of the minimum eigenvector of $\nabla^2 F(w_t)$. Thus, we have

$$\mathrm{Vol}(\mathcal{X}^{\epsilon}(w_t)) \leq \mathrm{Vol}(\mathbb{B}_0^{(d-1)}(\eta r))\cdot\eta r_0, \tag{40}$$

which gives us

$$\frac{\mathrm{Vol}(\mathcal{X}^{\epsilon}(w_t))}{\mathrm{Vol}(\mathbb{B}^d_{w_t}(\eta t)}) \leq \frac{\mathrm{Vol}(\mathbb{B}_0^{(d-1)}(\eta r))\cdot\eta r_0}{\mathrm{Vol}(\mathbb{B}^d_{w_t}(\eta t)}) = \frac{r_0}{r\sqrt{\pi}}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{r_0}{r\sqrt{\pi}}\sqrt{\frac{d+1}{2}} \leq 2\xi.$$

Hence, with probability at least $1-2\xi$, the perturbation lands in $\mathbb{B}^d_{w_t}(\eta t)\backslash\mathcal{X}^{\epsilon}(w_t)$. That is, with probability at least $1-\sqrt{\xi}$, the following holds

$$F(DPGD_{\epsilon}^{(\Gamma)}(w_t)) - F(w_t) \leq -\Phi.$$

Thus, we have the above inequality with probability at least $(1-\xi)(1-2\xi)(1-\sqrt{\xi}) \geq 1-3\sqrt{\xi}$. Reparametrizing $\xi' = 3\sqrt{\xi}$ only affects the factors in χ. □

Now, we prove Theorem 7.

Proof of Theorem 7. By Lemmas 8 and 9, we have, with probability at least $1-\frac{2B}{\Phi}\xi$, that the algorithm will find an α-SOSP in the following number of iterations

$$O(\frac{B}{\eta\alpha^2} + \frac{B\Gamma}{\Phi}) = O(\frac{B\chi^4}{\eta\alpha^2}).$$

What we need is that $\frac{\sqrt{2c_2\log\frac{1}{\delta}}TpL\log\frac{4T}{\xi}}{n\epsilon} \leq r = \alpha\chi^{-3}c^{-6}$, which means $n \geq \tilde{\Omega}(\frac{\sqrt{MB}\xi^5c^6\sqrt{2c_2\log\frac{1}{\delta}}pL}{\epsilon\alpha^2})$. Taking $\xi = \frac{2B}{\Phi}\xi$ only affects the log term.

This completes the proof. □

B.8. Proof of Theorem 8

Proof. By a similar argument given in the proof of Theorem 8, we known that there exist c_1, c_2 that make step 2 to step 6 $(\frac{\epsilon}{2}, \frac{\delta}{2})$-DP. Thus, the whole algorithm is (ϵ, δ)-DP.

By Lemma 13, we know that with probability at least $1 - \xi - \frac{T}{p^C}$

$$\|\epsilon_t\|_2 \leq \frac{\sqrt{c_2 \log \frac{1}{\delta} T d} L \log \frac{8T}{\xi}}{n\epsilon} = Err_1$$
$$\|H_t\|_2 \leq \frac{C\sqrt{c_3 \log \frac{1}{\delta} T} M d}{n\epsilon} = Err_2.$$

Now, we assume that the above event happens. By Theorem 7, we know that with probability at least $1 - \frac{\xi}{2}$, there exists $\|\nabla F(w_t)\|_2 \leq \frac{\alpha}{2}$ and $\lambda_{\min}(\nabla^2 F(w_t)) \geq -\sqrt{\frac{\rho\alpha}{2}}$. Thus, for this t, we have

$$\|g_t\|_2 \leq Err_1 + \frac{\alpha}{2} \leq \alpha$$
$$\lambda_{\min}(\tilde{H}_t) \geq \lambda_{\min}(\nabla^2 F(w_t)) - Err_2 \geq -\sqrt{\rho\alpha}.$$

These inequalities hold when $Err_1 \leq \frac{\alpha}{2}$ and $Err_2 \leq (1 - \sqrt{\frac{1}{2}})\sqrt{\rho\alpha}$. Thus, the size n should satisfy

$$n \geq \tilde{\Omega}\big(\max\{ \frac{\sqrt{BM \log \frac{1}{\delta} d} L \log \frac{1}{\xi}}{\epsilon\alpha^2}, \frac{\sqrt{\log \frac{1}{\delta} BM} M d \log \frac{1}{\xi}}{\rho\epsilon\alpha^2} \} \big).$$

Combining this with Theorem 7, we get the theorem. □

B.9. Proof of Theorem 9

Proof. First, we show the guarantee of (ϵ, δ)-DP. By Lemma 1, we know that $\sigma_1^2 = \frac{16 c_2 \log \frac{2}{\delta} L^2 T}{n^2 \epsilon^2}$, where c_2 is the constant in Lemma 1. Hence, it is $(\frac{\epsilon}{2}, \frac{\epsilon}{\delta})$-DP. Due to the L-smoothness, we have that for any pair of neighboring datasets D, D', $\|\nabla^2 \hat{L}(w, D) - \nabla^2 \hat{L}(w, D')\|_2 \leq 2L$, which means that $\|\nabla^2 \hat{L}(w, D) - \nabla^2 \hat{L}(w, D')\|_F \leq 2\sqrt{d} L$. This implies that if we view the Hessian matrix as a vector, the ℓ_2-sensitivity is $2\sqrt{d} L$. Also, due to symmetric structure, adding symmetric Gaussian matrix with each entry sampled from $\mathcal{N}(0, \sigma_2^2)$, where $\sigma_2^2 = \frac{c_3 T \log \frac{1}{\delta} M^2 d}{n^2 \epsilon^2}$, will ensure $(\frac{\epsilon}{2}, \frac{\delta}{2})$-DP. Thus, the algorithm is (ϵ, δ)-DP.

Then, we show the ability of escaping saddle points. For simplicity, we let $F(\cdot) = \hat{L}(\cdot, D)$, $\sqrt{\rho\alpha} = \gamma$, $\gamma' = \frac{\gamma}{2}$, $\alpha' = \frac{\alpha}{2}$, and $r = \frac{\Phi^2 \gamma'^2}{18\rho D^3}$.

We first show the following lemma by using the concentration of Gaussian distribution and the spectrum of symmetric Gaussian noise (Tao, 2012).

Lemma 13. *For any $0 < \xi < 1$, there exists a constant $C, c_3, c_2 > 0$ such that with probability at least $1 - \xi - \frac{T}{p^C}$, for any $t \in [T]$,*

$$\|\epsilon_t\|_2 \leq \frac{\sqrt{c_2 \log \frac{1}{\delta} T d} L \log \frac{8T}{\xi}}{n\epsilon} = Err_1 \tag{41}$$
$$\|H_t\|_2 \leq \frac{C\sqrt{c_3 \log \frac{1}{\delta} T} M d}{n\epsilon} = Err_2. \tag{42}$$

In the remaining analysis, we assume that the events in Lemma 13 happen, and the data size n is large enough such that

$$Err_1 \leq \min\{\frac{\Phi^2\gamma'^2}{18\rho D^4}, \frac{\alpha'}{4D}\} \tag{43}$$

$$Err_2 \leq \frac{\Phi\gamma'}{9D^2}. \tag{44}$$

Thus, n should be

$$n \geq \max\{\frac{18\sqrt{c_2 \log\frac{1}{\delta} Td}L\log\frac{8T}{\xi}\rho D^4}{\epsilon\Phi^2\gamma'^2}, \frac{4\sqrt{c_2\log\frac{1}{\delta}Td}L\log\frac{8T}{\xi}D}{\epsilon\alpha'}, \frac{9\sqrt{c_3\log\frac{1}{\delta}T}D^2Md}{\Phi\gamma'\epsilon}\}. \tag{45}$$

We now show the iteration complexity of Algorithm 5. First, we consider the case of $g_t^T(v_t - w_t) \leq -\alpha'$.

Lemma 14. *For w_t, if $g_t^T(v_t - w_t) \leq -\alpha'$, then we have*

$$F(w_{t+1}) \leq F(w_t) - \frac{\alpha'^2}{4D^2M}. \tag{46}$$

Proof of Lemma 14. By the M-smoothness of $F(\cdot)$, we have

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle\nabla F(w_t), w_{t+1} - w_t\rangle + \frac{M}{2}\|w_{t+1} - w_t\|_2^2 \\ &\leq F(w_t) + \eta\langle g_t, v_t - w_t\rangle + \eta\langle\epsilon_t, w_t - v_t\rangle + \frac{\eta^2 M D^2}{2} \\ &\leq F(w_t) - \eta\alpha' + \eta D Err_1 + \frac{\eta^2 M D^2}{2}. \end{aligned}$$

Taking $\eta = \frac{\alpha'}{D^2M}$, since $Err_1 \leq \frac{\alpha'}{4D}$, we have

$$F(w_{t+1}) \leq F(w_t) - \frac{\alpha'^2}{4D^2M}.$$

□

Lemma 15. *For a given w_t, if $g_t(v_t - w_t) \geq -\alpha'$ and $q(u_t) \leq -\Phi\gamma'$, then we have*

$$F(w_{t+1}) \leq F(w_t) - \frac{\Phi^3\gamma'^3}{6\rho^2D^6}. \tag{47}$$

Proof of Lemma 15.

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle\nabla F(w_t), w_{t+1} - w_t\rangle + \frac{1}{2}(w_{t+1} - w_t)^T\nabla^2 F(w_t)(w_{t+1} - w_t) + \frac{\rho}{6}\|w_{t+1} - w_t\|^2 \\ &\leq F(w_t) + \theta\langle\nabla F(w_t), u_t - w_t\rangle + \frac{\theta^2}{2}(u_t - w_t)^T\nabla^2 F(w_t)(u_t - w_t) + \frac{\theta^3\rho D^3}{6} \\ &\leq F(w_t) + \theta\langle g_t, u_t - w_t\rangle + \theta\langle\epsilon, w_t - u_t\rangle + \frac{\theta^2}{2}(u_t - w_t)^T\tilde{H}_t(u_t - w_t) - \frac{\theta^2}{2}(u_t - w_t)^T H_t(u_t - w_t) + \frac{\theta^3\rho D^3}{6} \\ &\leq F(w_t) + \theta r + \theta D Err_1 - \frac{\theta^2\Phi\gamma'}{2} + \frac{\theta^2 D^2 Err_2}{2} + \frac{\theta^3\rho D^3}{6}. \end{aligned}$$

Taking $\theta = \frac{\Phi\gamma'}{\rho D^3}$ and by the inequalities $Err_1 \leq \frac{\Phi^2\gamma'^2}{18D^4\rho}$, $Err_2 \leq \frac{\Phi\gamma'}{9D^2}$, and $r = \frac{\Phi^2\gamma'^2}{18\rho D^3}$, we get the lemma. □

By Lemmas 14 and 15, we know that under the events of Lemma 13, the algorithm terminates in $T = O(\max\{\frac{D^2MB}{\alpha'^2}, \frac{B\rho^2D^6}{\Phi^3\gamma'^3}\})$ iterations.

Next, we will show that under the events of Lemma 13, the output w_t is an (α, γ)-SOSP. From the theorem, we know that w_t satisfies the following conditions:

$$g_t^T(v - w_t) \geq -\alpha', \forall v \in \mathcal{C},$$
$$(u - w_t)^T \tilde{H}_t(u - w_t) \geq -\Phi\gamma', \forall u \in \mathcal{C}, g_t^T(u - w_t) \leq r.$$

We will first show that w_t satisfies the first order condition, that is

$$\max_{u\in C}\langle \nabla F(w_t), u - w_t\rangle \geq \min_{u\in C}(\langle g_t, u - w_t\rangle - DErr_1) \geq -\alpha' - DErr_1 \geq -\alpha.$$

We then show that w_t satisfies the second-order property.

Let $\mathcal{A} = \{w|\langle \nabla F(w_t), w - w_t\rangle = 0\}$ and $\mathcal{B} = \{w|g_t^t(w - w_t) \leq r\}$. We can show that $A \subseteq B$. This is due to the following. For any $w \in A$, $\langle \nabla F(w_t), w - w_t\rangle = 0$. Thus,

$$g_t^t(w - w_t) = \nabla F(w_t)^T(w - w_t) + \epsilon_t^T(w - w_t) \leq D \cdot Err_1 \leq r.$$

Finally, for any $w \in \mathcal{C}$, we have

$$\begin{aligned}(w - w_t)^T\nabla^2 F(w_t)(w - w_t) &= (w - w_t)^T H_t(w - w_t) - (w - w_t)^T\tilde{H}_t(w - w_t)\\ &\geq -\Phi\gamma' - D^2Err_2\\ \geq -\frac{10}{9}\Phi\frac{\gamma}{2} \geq -\frac{5}{9}\Phi\gamma &\geq -\gamma.\end{aligned}$$

Thus, for all $w \in C$ satisfying the condition of $\langle \nabla F(w_t), w - w_t\rangle = 0$, we have $(w - w_t)^T\nabla^2 F(w_t)(w - w_t) \geq -\gamma$.

Thus, n should satisfy

$$n \geq \tilde{\Omega}\Big(\max\{\frac{LD^7\sqrt{dMB\log\frac{1}{\delta}}\log\frac{1}{\xi}\rho^{1/4}}{\epsilon\Phi^{7/2}\alpha^2}, \frac{\sqrt{\log\frac{1}{\delta}dBM}LD^4\log\frac{1}{\xi}\rho^{1/4}}{\epsilon\alpha^2\Phi^{3/2}}, \frac{d\sqrt{BM^3\log\frac{1}{\delta}}D^5\log\frac{1}{\xi}}{\rho^{1/4}\Phi^{5/2}\alpha^{3/2}\epsilon}\}\Big).$$

□

References

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 464–473. IEEE, 2014.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, 2015.

Dalalyan, A. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pp. 678–689, 2017.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pp. 8759–8770, 2018.

Gilbarg, D. and Trudinger, N. S. *Elliptic partial differential equations of second order*. springer, 2015.

Pardoux, E. and Veretennikov, A. Y. On the poisson equation and diffusion approximation. i. *Annals of probability*, pp. 1061–1085, 2001.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

Tao, T. Topics in random matrix theory. *Graduate studies in Mathematics*, 132:46–47, 2012.

Veretennikov, A. Y. On polynomial mixing bounds for stochastic differential equations. *Stochastic processes and their applications*, 70(1):115–127, 1997.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Vollmer, S. J., Zygalakis, K. C., and Teh, Y. W. Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3126–3137, 2018.