Privacy-Enhanced Methods for the Predict-then-Optimize Framework

Sihan Wang
School of Mathematical Sciences, University of Science and Technology of China, wsh2021@mail.ustc.edu.cn

Hanzhang Qin
Department of Industrial Systems Engineering and Management and Institute of Operations Research and Analytics, National University of Singapore, Singapore, 117576, hzqin@nus.edu.sg

The "Smart Predict-then-Optimize" (SPO) framework has gained attention for its ability to integrate prediction and optimization in decision-making tasks. In this paper, we propose two novel gradient descent algorithms that incorporate the exponential mechanism and Gaussian mechanism within the SPO framework, aiming to strike a balance between data privacy and optimization performance. We provide a rigorous performance analysis of the algorithms, demonstrating their ability to preserve privacy without significantly affecting convergence rates. Moreover, we establish that under specific conditions, these algorithms achieve provably near-optimal accuracy in terms of the expected SPO loss. Experimental results further support our theoretical insights, highlighting the algorithms' robust performance in both private and non-private settings.

Key words: Differential Privacy, Smart Predict-then-Optimize, First-Order Method.

1. Introduction

In the era of big data and advanced machine learning, optimizing models and ensuring data privacy has become a top priority for businesses in the fields of healthcare (Dankar and El Emam 2013), finance (Li et al. 2019), and e-commerce (Reddy et al. 2023). Traditional gradient descent techniques for tuning the machine learning model parameters frequently face major difficulties in balancing sensitive data protection and optimization performance.

Differential privacy (Dwork 2006) has emerged as a robust framework for protecting individual data points during a machine learning process. By incorporating controlled noise into the machine learning procedure, differential privacy ensures that any data point's inclusion or exclusion has no significant impact on the learning outcome, protecting privacy without drastically reducing the model's performance.

Concurrently, decision-oriented learning has gained attention with specialized loss functions like the Smart Predict-then-Optimize (SPO) loss and its more practical version the SPO+ loss (Elmachtoub and Grigas 2022). These loss functions are designed to directly leverage the optimization

problem structure, measuring the decision error induced by certain predictions. SPO+ loss, in particular, offers tractable computational properties and near-optimal theoretical guarantees, making it a preferred choice in scenarios where prediction accuracy directly impacts optimization outcomes.

Despite significant advancements, existing data analytics approaches often struggle with the balance between maintaining strong privacy guarantees and achieving high optimization performance, especially when leveraging sophisticated loss functions like SPO+. Moreover, the extension of privacy-preserving techniques to generalized loss functions remains underexplored, limiting their applicability in broader optimization contexts.

In this paper, we address these challenges by proposing two novel gradient descent algorithms that ensure differential privacy in optimizing the SPO+ loss function. Our approach protects sensitive data while maintaining near-optimal convergence rates. We conduct a theoretical performance analysis of the proposed algorithms, demonstrating their efficacy in preserving privacy without substantial degradation in optimization performance. Furthermore, we extend our methodology to accommodate the nonconvex and thus more intractable SPO loss function under specific conditions, thereby broadening the algorithm's applicability across various optimization scenarios.

Private SPO optimizes decisions based on predicted outcomes while preserving data privacy, and addresses more complex decision-making problems typically applied in highly uncertain scenarios. On the contrary, Private convex optimization directly solves convex optimization problems with privacy constraints, focusing on the optimization of certain fully fixed objectives without integrating the prediction procedure with the decision-making process. As a result, the private SPO method can generate more reliable decisions in uncertain environments compared to the naive "first predict, and then optimize" method via the private convex optimization.

Our key contributions are summarized as follows:

1. We introduce two novel perturbed first-order methods and a smoothing scheme that integrates differential privacy with the SPO+ loss function. To the best of our knowledge, these are the first proposed algorithms for smart predict-then-optimize with provable privacy guarantees.

2. We prove that the proposed two algorithms both achieve $\mathcal{O}\left(\sqrt{\frac{1}{n}}+h(\varepsilon,\delta)\frac{1}{n}\right)$ SPO+ loss, and provide a matching lower bound (the function $h(\cdot)$ will be specified later).

3. We show the convergence analysis of the algorithms can also hold for the nonconvex SPO loss under certain conditions.

1.1. Related Literature

The smart "Predict-then-Optimize" framework was a recently proposed efficient approach for balancing the estimation error and optimization error in a general decision-making process (Elmachtoub and Grigas 2022, Bertsimas and Kallus 2020). There has been an enormous number of papers

to apply the SPO framework under different contexts, e.g., combinatorial optimization (Mandi et al. 2020), inventory management (Qi et al. 2023), vehicle routing (Soeffker et al. 2022), and maritime transportation Tian et al. (2023).

Differential privacy (DP) has been one of the most powerful concepts in terms of data privacy in machine learning. Since the inaugural works (Dwork 2006, Dwork et al. 2006), there have been numerous works that have further contributed to the theoretical development of DP, including performing mechanism designs to ensure DP (McSherry and Talwar 2007), the k-fold composition rule of DP (Dwork et al. 2010), the statistical framework of DP (Wasserman and Zhou 2010), private optimization algorithm design (Bassily et al. 2014), optimal rates for private convex optimization (Bassily et al. 2019).

Since the primary objective of this work is to introduce a novel framework for private SPO, our work can also be seen as a generalization from the more well-studied private convex optimization literature (see, e.g., Bassily et al. 2021, Gopi et al. 2022).

1.2. Notations

The dataset is denoted as $\mathcal{D} = \{(c_i, x_i)\}_{i=1}^n$, where $c_i \in \mathcal{R}^d$ is the cost function and $x_i \in \mathcal{R}^p$ is the feature vector. Let d represent the dimension of both the cost vector c and the decision vector w. In our algorithm, to keep it concise, we let $\ell(c, x) := \ell_{SPO+}(c, x)$, which is closed and convex. The loss function ℓ_μ is derived by smoothing the original loss function ℓ. For ℓ and ℓ_μ, we respectively define the expected risk and empirical risk: $\mathcal{L} = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}}[\ell(c, x)]$, $\mathcal{L}_\mu = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}}[\ell_\mu(c, x)]$ and $\hat{\mathcal{L}} = \frac{1}{n}\sum_{i=1}^n \ell(c_i, x_i)$, $\hat{\mathcal{L}}_\mu = \frac{1}{n}\sum_{i=1}^n \ell_\mu(c_i, x_i)$. L is the Lipschitz constant for l_μ. (ε, δ) are exclusive for Differential Privacy. $Proj(\cdot)$ is a projection function that maps a vector onto a certain set. In regret analysis, $\Omega(f(n))$ denotes the lower bound, meaning the regret grows at least as fast as $f(n)$, while $\mathcal{O}(f(n))$ denotes the upper bound, meaning the regret grows at most as fast as $f(n)$. Let S be the domain of decision w, which, in our experiments, is taken to be the unit sphere, i.e., $S = \{w \in \mathbb{R}^n : \|w\|_2 = 1\}$. $\mathbf{B}_w$ and $\mathbf{B}_B$ represent the absolute upper bound of w and B.

2. Preliminaries for Differential Privacy

Suppose that a data science team is developing a new health app, which aims to provide personalized health recommendations based on user data. To protect user privacy, they decided to implement differential privacy. In other words, even after seeing the app's recommendations for a specific user, it remains extremely difficult to infer or speculate about the individual's detailed personal information. The parameter ε in the following definition quantifies the degree of privacy protection.

DEFINITION 1. A randomized mechanism $\mathcal{M}$ is (ε, δ)-differentially private if for all datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ differing on at most one element (called neighboring datasets), and for all subsets S of the output space of $\mathcal{M}$, the following holds:

$$\mathbb{P}[\mathcal{M}(\mathcal{D}_1) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(\mathcal{D}_2) \in S] + \delta$$

Additionally, it is called ε-DP when $\delta = 0$.

A lower ε value indicates stronger privacy guarantees, as it limits the amount of information that can be inferred about any individual from the data analysis results. Conversely, a higher ε value allows for more accurate analysis but provides less privacy protection. This parameter allows data scientists and privacy experts to precisely tune the trade-off between data utility and individual privacy in their applications.

The Exponential mechanism and the Gaussian mechanism are both effective methods used in the field of differential privacy to ensure that the output of an algorithm doesn't reveal too much information about any individual's data. The Gaussian mechanism adds noise drawn from a Gaussian distribution directly to the numeric output, while the exponential mechanism introduces randomness through a probabilistic selection based on a utility function rather than adding noise directly to the output.

2.1. Exponential Mechanism

DEFINITION 2. The Exponential Mechanism $\mathcal{M}$ selects an output r with probability:

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) = r] = \frac{\exp(\varepsilon \cdot u(\mathcal{D}, r))}{\sum_{r' \in \mathcal{R}} \exp(\varepsilon \cdot u(\mathcal{D}, r'))}$$

where $u(\mathcal{D}, r)$ is a utility function, Δu is the sensitivity of the utility function u, i.e. for any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$,$\Delta u = \max_{r \in \mathcal{R}} |u(\mathcal{D}, r) - u(\mathcal{D}', r)|$.

THEOREM 1 **(McSherry and Talwar (2007))**. *$\mathcal{M}$ is $2\Delta u \cdot \varepsilon$-differential private.*

Throughout this paper, we consider the utility function u as a binary function taking values only in 0 or 1, thus $\Delta u = 1$. Meanwhile, we adopt a uniform sampling scheme:

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \sim \mathcal{P}_1] = \frac{e^{\varepsilon}}{e^{\varepsilon} + 1}, \ \mathbb{P}[\mathcal{M}(\mathcal{D}) \sim \mathcal{P}_2] = \frac{1}{e^{\varepsilon} + 1}$$

where $\mathcal{P}_1, \mathcal{P}_2$ are respectively uniform distribution over disjoint regions.

2.2. Gaussian Mechanism

DEFINITION 3. Given a query function $f : \mathcal{D} \to \mathbb{R}^k$ and a dataset $\mathcal{D}$, the Gaussian Mechanism outputs:

$$\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + \mathcal{N}(0, \sigma^2)$$

where $\mathcal{N}(0, \sigma^2)$ is noise drawn from a Gaussian distribution with mean 0 and standard deviation σ.

From Dwork et al. (2014) we obtain the following statement:

THEOREM 2. *Let $\varepsilon \in (0,1)$ be arbitrary. For $c^2 > 2\ln(1.25/\delta)$, the Gaussian Mechanism with parameter $\sigma \geq c\Delta_2(f)/\varepsilon$ is (ε,δ)-differentially private, where $\Delta_2(f) = \max_{\|x-y\|_1=1} \|f(x)-f(y)\|_2$.*

2.3. Post-Processing

Suppose $\mathcal{M}$ is a (ε,δ)-DP mechanism. It is evident that any post-processing operation applied to the output of $\mathcal{M}$, provided it does not introduce new data-dependent information, preserves the (ε,δ)-differential privacy guarantee of the overall process. In other words, f is any data-independent function, considering the pre-image f^{-1} we have $\mathbb{P}[f\circ\mathcal{M}(\mathcal{D})\in S'] = \mathbb{P}[\mathcal{M}(\mathcal{D})\in S], \quad \mathbb{P}[f\circ\mathcal{M}(\mathcal{D}')\in S'] = \mathbb{P}[\mathcal{M}(\mathcal{D}')\in S]$ where $S = f^{-1}(S')$. Therefore $\mathbb{P}[f\circ\mathcal{M}(\mathcal{D})\in S'] \leq e^{\varepsilon}\mathbb{P}[f\circ\mathcal{M}(\mathcal{D}')\in S'] + \delta$, i.e. $f\circ\mathcal{M}$ is also (ε,δ)-DP. The post-processing property is straightforward to understand: if a process already satisfies a certain level of differential privacy, any subsequent post-processing of its results will not compromise the existing privacy guarantees.

3. Preliminaries for SPO

Before introducing the SPO framework, let us describe the predict-then-optimize framework. Namely, predict-then-optimize involves making predictions and optimizing based on those predictions, where the predictions always depend on the current features. Mathematically, this can be expressed as solving the contextual stochastic optimization problem:

$$\min_{w\in S} \mathbb{E}_{c\sim\mathcal{D}_x}[c^\top w \mid x] = \min_{w\in S} \mathbb{E}_{c\sim\mathcal{D}_x}[c \mid x]^\top w,$$

where c is the predicted vector based on current feature x, w is the decision vector based on the prediction c, $\mathcal{D}_x$ is the conditional distribution of c given x. So, how can we make the predict-then-optimize framework "smart"?

Traditional loss functions typically focus on the deviation of prediction, for example, mean squared error and hinge loss. However, the actual "loss" remains constant as long as the decision w is the same despite the prediction c may be different as shown in Figure 1.

3.1. Smart Predict-then-Optimize

Following Elmachtoub and Grigas (2022), first we define the *nominal (downstream) optimization problem*, which is of the form

$$\begin{aligned} P(c): \quad z^*(c) := \min_{w}\ & c^T w \\ \text{s.t.}\quad & w \in S, \end{aligned}$$

where $w\in\mathbb{R}^d$ are the decision variables, $c\in\mathbb{R}^d$ is the problem data describing the linear objective function, and $S\subseteq\mathbb{R}^d$ is a nonempty, compact (i.e., closed and bounded), and convex set representing the feasible region.

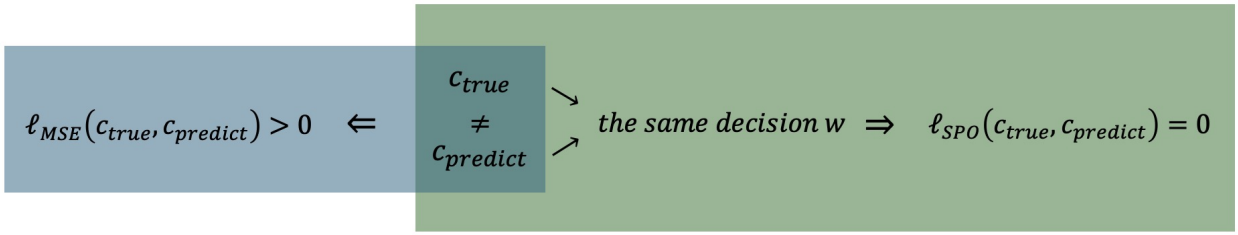

Figure 1 **This image visually illustrates the difference between the SPO loss function and traditional loss functions, such as Mean Squared Error (MSE) shown in the diagram:** $\ell_{MSE}(c_{true}, c_{predict}) = \|c_{true} - c_{predict}\|_2$.

Definition 4 (Elmachtoub and Grigas (2022)). Given a cost vector prediction $\hat{c}$ and a realized cost vector c, the SPO loss $\ell_{SPO}(\hat{c}, c)$ is defined as

$$\ell_{SPO}(\hat{c}, c) := \max_{w \in W^*(\hat{c})} \{c^T w\} - z^*(c).$$

To begin the derivation of the SPO+ loss, we first observe that for any $\alpha \in \mathbb{R}$, the SPO loss can be written as

$$\ell_{SPO}(\hat{c}, c) = \max_{w \in W^*(\hat{c})} \{c^T w - \alpha \hat{c}^T w\} + \alpha z^*(\hat{c}) - z^*(c), \tag{1}$$

since $z^*(\hat{c}) = \hat{c}^T w$ for all $w \in W^*(\hat{c})$. Clearly, replacing the constraint $w \in W^*(\hat{c})$ with $w \in S$ in (1) results in an upper bound. Since this is true for all values of α, then

$$\ell_{SPO}(\hat{c}, c) \leq \inf_{\alpha} \left\{ \max_{w \in S} \{c^T w - \alpha \hat{c}^T w\} + \alpha z^*(\hat{c}) \right\} - z^*(c). \tag{2}$$

Setting $\alpha = 2$, we obtain the SPO+ loss as follows:

Definition 5 (Elmachtoub and Grigas (2022)). Given a cost vector prediction $\hat{c}$ and a realized cost vector c, the SPO+ loss is defined as

$$\ell_{SPO+}(\hat{c}, c) := \max_{w \in S} \{c^T w - 2\hat{c}^T w\} + 2\hat{c}^T w^*(c) - z^*(c).$$

Noticing $z^*(\hat{c}) \leq \hat{c}^T w^*(c)$, we know that $\ell_{SPO} \leq \ell_{SPO+}$ directly, which will be restate in Section 6 "From SPO+ to SPO".

Throughout the paper, we mainly focus on SPO+ loss since it is convex and thus can be written as a conjugate of another function, i.e. there exists a function g such that $f(x) = g^*(x) = \sup_y \{\langle x, c \rangle - g(y)\}$.

$$\begin{aligned}
\ell_{SPO+}(c, \hat{c}) &:= \max_{w \in S} \left\{ c^T w - 2\hat{c}^T w \right\} + 2\hat{c}^T w^*(c) - z^*(c) \\
&= \max_{w \in S} \left\{ c^T w - 2\hat{c}^T w + 2\hat{c}^T w^*(c) - z^*(c) \right\} \\
&= \max_{w \in S} \left\{ \langle -2w + 2w^*(c), \hat{c} \rangle - \langle -w + w^*(c), c \rangle \right\}
\end{aligned}$$

To clarify, we restate the definition of DP through the lens of SPO:

DEFINITION 6. A randomized algorithm $\mathcal{A}$ is (ε, δ)-DP if $\forall$ neighboring datasets $\mathcal{D}_1 = \{(c_1, x_1), ..., (c_i, x_i), ..., (c_n, x_n)\}, \mathcal{D}_2 = \{(c_1, x_1), ..., (c_i', x_i'), ..., (c_n, x_n)\}$ and $\forall\ w \in S$, the following holds:

$$\mathbb{P}[\mathcal{A}(\mathcal{D}_1) = w] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{A}(\mathcal{D}_2) = w] + \delta.$$

3.2. Smoothing

Given that the SPO+ loss function is non-smooth, we apply smoothing techniques to approximate its gradient for faster optimization rates. Let us first define smoothness formally.

DEFINITION 7. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be Lipschitz continuous if there exists a constant $K \geq 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq K\|\mathbf{x} - \mathbf{y}\|,$$

DEFINITION 8. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called β-smooth if it is differentiable and its gradient is Lipschitz continuous with constant $\beta > 0$. This means that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta\|\mathbf{x} - \mathbf{y}\|.$$

Note that, if further, the domain is bounded, then $\|f(\mathbf{x}) - f(\mathbf{y})\|$ is also bounded by $\|\mathbf{x} - \mathbf{y}\|$.

LEMMA 1. *If $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth and the domain is bounded, then f is also Lipschitz continuous.*

We use the smoothing technique introduced in Chen (2020):

Suppose $f = g^*$, namely,

$$f(\mathbf{x}) = \sup_{\mathbf{z}}\{\langle \mathbf{z}, \mathbf{x}\rangle - g(\mathbf{z})\}.$$

We can build a smooth approximation of f by adding a strongly convex component to its dual, namely,

$$f_\mu(\mathbf{x}) = \sup_{\mathbf{z}}\{\langle \mathbf{z}, \mathbf{x}\rangle - g(\mathbf{z}) - \mu d(\mathbf{z})\} = (g + \mu d)^*(\mathbf{x}),$$

for some 1-strongly convex and continuous function $d(\cdot) \geq 0$. We have the following two properties:

1. $g + \mu d$ is μ-strongly convex $\implies$ f_μ is $\frac{1}{\mu}$-smooth
2. $f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D$ with $D := \sup_x d(x)$

More specifically, for the SPO+ loss function, the smoothed function with respect to the prediction variable $\hat{c}$ is given by

$$\ell_\mu(c, \hat{c}) = \max_{w \in S}\left\{\langle -2w + 2w^*(c), \hat{c}\rangle - \langle -w + w^*(c), c\rangle - \frac{1}{2}\mu\|w\|_2^2\right\}$$

4. The Exponential Descent Method

We develop a novel exponential mechanism for optimizing the SPO loss, called exponential descent. This mechanism is inspired by Gopi et al. (2022).

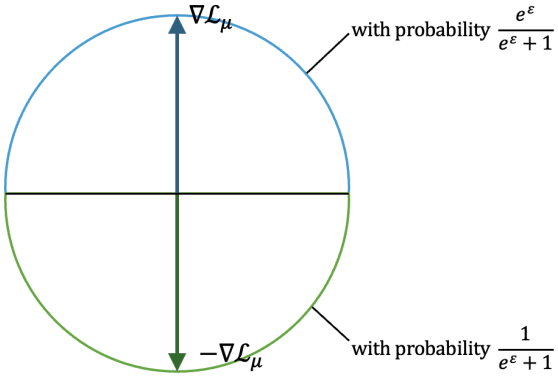

Figure 2 **The descent direction in Algorithm 1.**

Recall the SPO+ Loss Function:

$$l_{SPO+}(c,\hat{c}) := \max_{w\in S}\left\{c^T w - 2\hat{c}^T w\right\} + 2\hat{c}^T w^*(c) - z^*(c)$$

where $\hat{c} = \hat{B}x$, $\hat{B} = (\hat{B}^1, ..., \hat{B}^d)^T$ is unknown, d is the dimension of $\hat{c}$, $\hat{B}^j$ is a p-dimension vector. $\hat{c}^T w = \sum_{j=1}^d (x^T \hat{B}^j) w_j = \sum_{j=1}^d \sum_{i=1}^p (x_i \hat{B}_i^j) w_j$, we can consider $\hat{B}$ as an pd-dimension vector. We define $B := \mathrm{vec}(\hat{B})$ as the vector obtained by vectorizing matrix B.

Our concern lies in the potential for inferring the original data pair (c, x) given knowledge of the decision w. Consequently, we conduct an analysis of the algorithm's differential privacy characteristics with respect to the decision variable w.

LEMMA 2. *For each iteration, ED is ε-Differential Private.*

The post-processing property of differential privacy ensures that our algorithm maintains ε-differential privacy when only the final decision w is disclosed, as each step of the algorithm independently satisfies ε-differential privacy. However, a more complex scenario arises when considering the potential disclosure of decisions w generated at each intermediate step of the algorithm.

To address this scenario, we now turn our attention to defining the composition of differential privacy, which will allow us to analyze the cumulative privacy guarantees when multiple outputs from the privacy-preserving process are revealed. Further details and concepts related to composition can be found in the appendix. By Corollary 8.3.3 in Duchi (2024) we know that:

Algorithm 1: Exponential Descent (ED)

Input: Dataset $\mathcal{D}=\{(x_i,c_i)\}_{i=1}^n \in \mathcal{X}^n \times \mathcal{C}^n$; an upcoming feature vector $\mathbf{x}_{n+1}$

Set initial point $B_1=0$;

for $t=1$ *to* $T-1$ **do**

Draw a descending direction;

$$G_t|\nabla\hat{\mathcal{L}}_\mu(B_t;\mathcal{D}) \sim \begin{cases} \text{Uniform}(\{\|v\|_2=1 : \langle v, \nabla\hat{\mathcal{L}}_\mu(B_t;\mathcal{D})\rangle \geq 0\}) \text{ w.p. } \frac{e^\varepsilon}{e^\varepsilon+1} \\ \text{Uniform}(\{\|v\|_2=1 : \langle v, \nabla\hat{\mathcal{L}}_\mu(B_t;\mathcal{D})\rangle \leq 0\}) \text{ w.p. } \frac{1}{e^\varepsilon+1} \end{cases}$$

Update the parameter;

$$B_{t+1} \leftarrow B_t - \eta\|\nabla\hat{\mathcal{L}}_\mu(B_t;\mathcal{D})\|_2 G_t$$

end

Set final parameter:

$$B \leftarrow \text{Proj}\left(\frac{1}{T}\sum_{t=1}^T B_t\right)$$

Output: $\hat{c}=\hat{B}\mathbf{x}_{n+1}$ and $w^*(\hat{c})$

LEMMA 3. *Assume that each channel is ε-differentially private. Then the composition of k such channels is $k\varepsilon$-differentially private. Additionally, the composition of k such channels is*

$$\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k\log\frac{1}{\delta}}\cdot\varepsilon, \delta\right)$$

differentially private for all $\delta>0$.

Let each iteration satisfies $\varepsilon_0 = \frac{\sqrt{2\log\frac{1}{\delta}+2\varepsilon}-\sqrt{2\log\frac{1}{\delta}}}{\sqrt{3k}}$-Differential Privacy, then we have:

THEOREM 3. *For any number of iterations $T\geq 1$, Algorithm ED is (ε,δ)-DP.*

Next, we analyze the regret of the algorithm. Following the approach in Chen and Chua (2023), we can decompose $\mathcal{R}(\mathbb{P}) = \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}[\mathcal{L}(B)] - \mathcal{L}(B^*)$ into three parts: $\mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\mathcal{L}(B,\mathcal{D}) - \hat{\mathcal{L}}(B,\mathcal{D})\right] + \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\hat{\mathcal{L}}(B) - \hat{\mathcal{L}}(B^*)\right] + \left[\hat{\mathcal{L}}(B^*) - \mathcal{L}(B^*)\right]$, then derive upper bound bound for each of them. Through this process, we find that $\mathcal{R}(\mathbb{P})$ is of the form $a\eta + b\frac{1}{\eta} \leq 2\sqrt{ab}$ where η is the step size.

THEOREM 4 **(Upper Bound)**. *Take*

$$\eta = \frac{\sqrt{\frac{e^\varepsilon-1}{e^\varepsilon+1}\frac{B_B^2}{4T}}}{\sqrt{\frac{L^2(1+T)}{n} + \frac{e^\varepsilon-1}{e^\varepsilon+1}(\frac{L^2}{4}+\frac{B_w^2}{4})}}, \quad T=n^2,$$

then

$$\mathcal{R}(\mathbb{P}) = \mathcal{O}\sqrt{\frac{1}{n}} + \left(\frac{e^\varepsilon-1}{e^\varepsilon+1}\frac{1}{n}\right)$$

where $\mathcal{R}(\mathbb{P}) := \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}[\mathcal{L}(B) - \mathcal{L}(B^*)]$ *is the regret of Algorithm GP under distribution* $\mathbb{P}$.

In terms of the lower bound, from Bassily et al. (2019) we know that the lower bound of $\Omega(1/\sqrt{n})$ for the excess loss of non-private stochastic convex optimization (SCO) can be applied for private SCO. Further we show in the proof of Theorem 6.3 that $\mathcal{R}(\mathbb{P})$ has a lower bound of the form $\Omega(\frac{e^\varepsilon-1}{e^\varepsilon+1}\frac{1}{n})$.

THEOREM 5 **(Lower Bound)**. *Let* $n \in \mathbb{N}$ *and* $\delta > 0$. *There exists a dataset* $\mathcal{D} = \{(c_i, x_i)\}_{i=1}^n$ *with probability at least* $\frac{1}{2}$ *such that:*

$$\mathcal{R}(\mathbb{P}) = \Omega\left(\frac{1}{\sqrt{n}} + \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\frac{1}{n}\right).$$

5. The Gaussian Descent Method

Algorithm 2: Gaussian Descent (GD)

Input: Dataset $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^n \in \mathcal{X}^n \times \mathcal{C}^n$; an upcoming feature vector $\mathbf{x}_{n+1}$

Set initial point $B_1 = 0$;

for $t = 1$ *to* $T-1$ **do**

Draw a descending direction with Gaussian noise;

$$G_t \sim \nabla\hat{\mathcal{L}}_\mu(B_t; \mathcal{D}) + \mathcal{N}(0, \sigma^2 I_d)$$

where $\sigma^2 = \frac{8L^2T\ln(1/\delta)}{n^2\varepsilon^2}$;

Update the parameter;

$$B_{t+1} \leftarrow B_t - \eta G_t$$

end

Set final parameter:

$$B \leftarrow \text{Proj}\left(\frac{1}{T}\sum_{t=1}^{T} B_t\right)$$

Output: $\hat{c} = \hat{B}\mathbf{x}_{n+1}$ and $w^*(\hat{c})$

Now we turn to another famous mechanism from the differential privacy literature: the Gaussian mechanism. We call our corresponding algorithm the Gaussian descent method. Considering that the Gaussian mechanism satisfies (ε, δ)-differential privacy, we have:

THEOREM 6. *For any number of iterations* $T \geq 1$, *Algorithm GD is* (ε, δ)*-DP.*

In fact, due to the nature of differential privacy being independent of the loss function, we can directly see the above theorem from Theorem 3 in Chen and Chua (2023), as the two theorems are essentially identical. Similarly, we can derive similar results compared to those of the ED method:

THEOREM 7 **(Upper Bound)**. *Take*

$$\eta = \frac{\sqrt{\frac{\|\mathbf{B}_B\|^2}{2T}}}{\sqrt{\frac{L^2}{2} + \frac{d\sigma^2}{2} + \frac{B_w^2}{4} + \frac{L^2(1+T)}{n}}}, \quad T = n^2,$$

then

$$\mathcal{R}(\mathbb{P}) = \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right).$$

THEOREM 8 **(Lower Bound)**. *Let $n \in \mathbb{N}$ and $\delta > 0$. There exists a dataset $D = \{(c_i, x_i)\}_{i=1}^n$ with probability at least $\frac{1}{2}$ such that:*

$$\mathcal{R}(\mathbb{P}) = \Omega\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right).$$

6. From SPO+ to SPO

In this section, we will extend the convergence results for the SPO+ loss to the SPO loss. From Proposition 3 in Elmachtoub and Grigas (2022) we have:

LEMMA 4. *For every pair $(c, \hat{c})$, $\ell_{SPO+}(c, \hat{c}) \geq \ell_{SPO}(c, \hat{c})$.*

Next, we present that in terms of mathematical expectation, the SPO loss and the SPO+ loss take the same value at $\mathbb{E}[c|x]$.

LEMMA 5. *When $c \leq 2\mathbb{E}[c|x]$ almost surely, we have*

$$\begin{aligned}\mathbb{E}_{c,x}\left[\max_{w\in S}\{c^T w - 2\mathbb{E}[c|x]^T w\}\right] &= \mathbb{E}_x\left[\max_{w\in S}\{\mathbb{E}[c|x]^T w - 2\mathbb{E}[c|x]^T w\}\right] \\ &= \mathbb{E}_x\left[\max_{w\in S}\{-\mathbb{E}[c|x]^T w\}\right].\end{aligned}$$

THEOREM 9. *If $W^*(\mathbb{E}[c|x])$ is a singleton a.s. and $2\mathbb{E}[c|x] \geq c \geq 0$ a.s. then*

$$\mathbb{E}_{c,x}[\ell_{SPO+}(\hat{c} = \mathbb{E}[c|x], c)] = \mathbb{E}_{c,x}[\ell_{SPO}(\hat{c} = \mathbb{E}[c|x], c)].$$

Why do we care about the relationship between the two losses at $\mathbb{E}[c|x]$? This is because from the proof of Theorem 1 in Elmachtoub and Grigas (2022), we observe that $\mathbb{E}[c|x]$ minimizes both the SPO and SPO+ risk, if the following four conditions hold:

1. Almost surely, $W^*(\mathbb{E}[c|x])$ is a singleton, i.e., $\mathbb{P}_x(|W^*(\mathbb{E}[c|x])| = 1) = 1$.
2. For all $x \in \mathcal{X}$, the distribution of $c|x$ is centrally symmetric about its mean $\mathbb{E}[c|x]$.

3. For all $x \in \mathcal{X}$, the distribution of $c|x$ is continuous on all of $\mathbb{R}^d$.

4. The interior of the feasible region S is nonempty.

Therefore, we have the following corollary given the assumptions stated above and the conditions of Theorem 9 (namely, $2\mathbb{E}[c|x] \geq c \geq 0$):

COROLLARY 1. $\mathbb{E}_{c,x}\left[\ell_{SPO+}(\hat{B}x) - \ell_{SPO+}(\mathbb{E}[c|x])\right] \geq \mathbb{E}_{c,x}\left[\ell_{SPO}(\hat{B}x) - \ell_{SPO}(\mathbb{E}[c|x])\right]$.

As a special case, when there exists a matrix B^* such that $\mathbb{E}[c|x] = B^*x$ and $2\mathbb{E}[c|x] \geq c \geq 0$, we can infer convergence in the SPO framework from convergence in the SPO+ framework:

$$\begin{aligned}
&\mathbb{E}_{c,x}\left[\ell_{SPO}(\hat{B}x) - \ell_{SPO}(B^*x)\right] \\
&= \mathbb{E}_{c,x}\left[\ell_{SPO}(\hat{B}x) - \ell_{SPO}(\mathbb{E}[c|x])\right] \\
&\leq \mathbb{E}_{c,x}\left[\ell_{SPO+}(\hat{B}x) - \ell_{SPO+}(\mathbb{E}[c|x])\right] \\
&= \mathcal{O}\left(\sqrt{\frac{1}{n}} + h(\varepsilon,\delta)\frac{1}{n}\right),
\end{aligned}$$

where $h(\varepsilon,\delta) = \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$ for the Exponential Mechanism and $h(\varepsilon,\delta) = \sqrt{d\ln(1/\delta)}$ for the Gaussian Mechanism.

7. Numerical Experiments

To validate the effectiveness of the proposed algorithms and embody the impact of the (ε,δ) parameters on convergence, we conducted several experiments as follows. We simply assume that the real dataset satisfies $c = \hat{B}x$, where $\hat{B} = I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathcal{R}^{3\times 3}$, $c, x \in \mathcal{R}^3$. For each sample size n, the number of iterations is set to n^2, as stated in the previous theorem. Additionally, for each n, we randomly generate n values of x, and then obtain n corresponding values of c. First, we consider the comparison of SPO loss convergence between two algorithms (exponential mechanism and Gaussian mechanism) and the non-private case under the setting $(\varepsilon,\delta) = (10, 0.4)$ in Figure 3.

Recall the definition of Differential Privacy: $\mathbb{P}[\mathcal{M}(\mathcal{D}_1) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(\mathcal{D}_2) \in S] + \delta$. For different parameter pairs (ε,δ), based on the definition of DP, we can infer that the convergence rate decreases as ε increases. For δ, which means with probability of $1-\delta$ achieving DP, we can deduce that as δ increases, the DP guarantee becomes weaker, while the convergence improves. We show the SPO loss of the Exponential mechanism and Gaussian mechanism with different sample sizes and parameter pairs (ε,δ). We show the convergence of exponential mechanism and Gaussian mechanism with different sample sizes and parameter pairs (ε,δ).

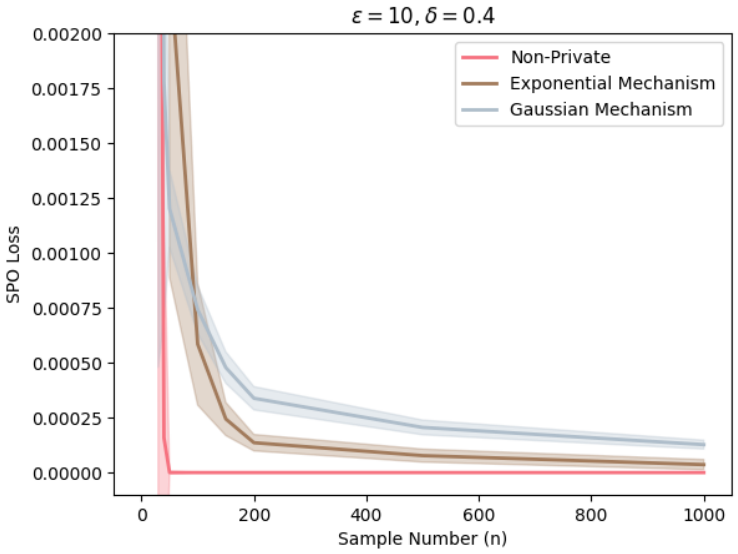

Figure 3 **The figure shows the SPO loss curves as a function of the sample number n for three cases: non-private, exponential mechanism and Gaussian mechanism. Due to inherent randomness, we repeated each experiment 5 times for each n. The solid lines represent the mean values, while the shaded regions reflect the variance across the 5 repetitions.**

	Exponential Mechanism		Gaussian Mechanism	
epsilon	Mean	Variance	Mean	Variance
0.1	0.0102	3.39e-05	0.0106	4.9e-05
1	6.85e-05	1.41e-09	0.00222	1.73e-06
5	1.99e-05	1.12e-10	0.000233	4.77e-08
10	2.19e-05	4.5e-10	0.000148	3.92e-09

Gaussian Mechanism		
delta	Mean	Variance
0.1	0.000281	1.97e-08
0.25	0.000116	6.24e-09
0.4	0.000171	4.95e-09

Table 1 **Comparison of the Exponential and Gaussian Mechanisms for $n = 500$. The table on the left shows the mean and variance for both mechanisms under different values of ϵ, illustrating their approximate convergence behavior. The table on the right shows the mean and variance of the Gaussian mechanism under different values of δ, where smaller variances indicate tighter convergence to the expected output.**

8. Conclusion

In this paper, we have proposed two first-order methods for smart predict-then-optimize with privacy guarantees. We have derived minimax optimal rates for the methods, by analyzing the convergence rates of the respective optimization algorithms and providing hard instances with matching complexity lower bounds. Our numerical experiments further validated the efficacy of our proposed methods. Future directions include studying algorithms with stronger guarantees for the SPO loss (we have primarily focused on the SPO+ loss), developing optimization algorithms for the multi-stage SPO/SPO+ loss, and ensuring data privacy via SPO in real-world applications.

References

Bassily, Raef, Vitaly Feldman, Kunal Talwar, Abhradeep Guha Thakurta. 2019. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems* **32**.

Bassily, Raef, Cristóbal Guzmán, Michael Menart. 2021. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems* **34** 9317–9329.

Bassily, Raef, Adam Smith, Abhradeep Thakurta. 2014. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv e-prints* arXiv–1405.

Bertsimas, Dimitris, Nathan Kallus. 2020. From predictive to prescriptive analytics. *Management Science* **66**(3) 1025–1044.

Bousquet, Olivier, André Elisseeff. 2002. Stability and generalization. *The Journal of Machine Learning Research* **2** 499–526.

Chen, Du, Geoffrey A Chua. 2023. An algorithmic approach to managing supply chain data security: The differentially private newsvendor. *Nanyang Business School Research Paper* (23-22).

Chen, Yuxin. 2020. Smoothing for nonsmooth optimization.

Dankar, Fida Kamal, Khaled El Emam. 2013. Practicing differential privacy in health care: A review. *Trans. Data Priv.* **6**(1) 35–67.

Duchi, John. 2024. *Lecture Notes for Statistics 311/Electrical Engineering 377*. Stanford.

Dwork, Cynthia. 2006. Differential privacy. *International colloquium on automata, languages, and programming*. Springer, 1–12.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

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

Dwork, Cynthia, Guy N Rothblum, Salil Vadhan. 2010. Boosting and differential privacy. *2010 IEEE 51st annual symposium on foundations of computer science*. IEEE, 51–60.

Elmachtoub, Adam N, Paul Grigas. 2022. Smart "predict, then optimize". *Management Science* **68**(1) 9–26.

Gopi, Sivakanth, Yin Tat Lee, Daogao Liu. 2022. Private convex optimization via exponential mechanism. *Conference on Learning Theory*. PMLR, 1948–1989.

Hardt, Moritz, Ben Recht, Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. *International conference on machine learning*. PMLR, 1225–1234.

Li, Xinyi, Yinchuan Li, Hongyang Yang, Liuqing Yang, Xiao-Yang Liu. 2019. Dp-lstm: Differential privacy-inspired lstm for stock prediction using financial news. *arXiv preprint arXiv:1912.10806* .

Mandi, Jayanta, Peter J Stuckey, Tias Guns, et al. 2020. Smart predict-and-optimize for hard combinatorial optimization problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. 1603–1610.

McSherry, Frank, Kunal Talwar. 2007. Mechanism design via differential privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 94–103.

Qi, Meng, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu, Zuo-Jun Shen. 2023. A practical end-to-end inventory management model with deep learning. *Management Science* **69**(2) 759–773.

Reddy, Vijay Mallik, et al. 2023. Data privacy and security in e-commerce: Modern database solutions. *International Journal of Advanced Engineering Technologies and Innovations* **1**(03) 248–263.

Shalev-Shwartz, Shai, Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Soeffker, Ninja, Marlin W Ulmer, Dirk C Mattfeld. 2022. Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review. *European Journal of Operational Research* **298**(3) 801–820.

Tian, Xuecheng, Ran Yan, Yannick Liu, Shuaian Wang. 2023. A smart predict-then-optimize method for targeted and cost-effective maritime transportation. *Transportation Research Part B: Methodological* **172** 32–52.

Wasserman, Larry, Shuheng Zhou. 2010. A statistical framework for differential privacy. *Journal of the American Statistical Association* **105**(489) 375–389.

A. Proof of Lemma 2

Proof of Lemma 2

$$\begin{aligned} l_{\text{SPO+}}(c,\hat{c}) &:= \max_{w\in S}\left\{c^T w - 2\hat{c}^T w\right\} + 2\hat{c}^T w^*(c) - z^*(c) \\ &= \max_{w\in S}\left\{c^T w - 2\hat{c}^T w + 2\hat{c}^T w^*(c) - z^*(c)\right\} \\ &= \max_{w\in S}\left\{\langle -2w + 2w^*(c), \hat{c}\rangle - \langle -w + w^*(c), c\rangle\right\} \end{aligned}$$

$$l_\mu(c,\hat{c}) = \max_{w\in S}\left\{\langle -2w + 2w^*(c), \hat{c}\rangle - \langle -w + w^*(c), c\rangle - \frac{1}{2}\mu\|w\|_2\right\} := \max_{w\in S} f(w,c,\hat{c})$$

$$\frac{\partial f}{\partial w} = -2\hat{c} + c - \mu w = 0 \iff w = \frac{1}{\mu}(-2\hat{c} + c)$$

The projection of w onto the set S, w_s is continuous and unique with respect to w when S is convex. What's more, w is unique with respect to $\hat{c}$, $\hat{c}$ is unique with respect to the descent direction G_t. Let w_t, w_t' and G_t, G_t' respectively be decision made in l_μ and randomized descent direction trained on neighboring dataset $\mathcal{D}, \mathcal{D}'$in t th iteration.

$$\frac{\mathbb{P}(w_t = W)}{\mathbb{P}(w_t' = W)} = \frac{\mathbb{P}(G_t = z)}{\mathbb{P}(G_t' = z)} \le \frac{e^\varepsilon}{e^\varepsilon + 1} \Big/ \frac{1}{e^\varepsilon + 1} = e^\varepsilon$$

□

B. The Composition of DP Mechanisms

The following definitions are from Section 3 in Dwork et al. 2010. Let $\mathbb{M}$ be a family of database access mechanisms. (For example $\mathbb{M}$ could be the set of all ε-differentially private mechanisms.) For a probabilistic adversary A, we consider two experiments, Experiment 0 and Experiment 1, defined as follows.

Definition 9 (k-fold Composition Experiment b for mechanism family $\mathbb{M}$ and adversary A). For $i = 1, \ldots, k$:

1. A outputs two adjacent databases x_i^0 and x_i^1, a mechanism $\mathcal{M}_i \in \mathbb{M}$, and parameters w_i.
2. A receives $y_i \leftarrow \mathcal{M}_i(w_i, x_{i,b})$.

We allow the adversary A above to be stateful throughout the experiment, and thus it may choose the databases, mechanisms, and the parameters adaptively depending on the outputs of previous mechanisms.

Definition 10 (Max-divergence and Approximate Max-divergence). For probability distributions P and Q defined over a set Ω, we define:

1. The max-divergence $D_\infty(P\|Q)$ as:

$$D_\infty(P\|Q) = \max_{S\subseteq\Omega} \ln \frac{P(S)}{Q(S)}$$

2. The δ-approximate max-divergence $D_\infty^\delta(P\|Q)$ as:

$$D_\infty^\delta(P\|Q) = \max_{S\subseteq\Omega: P(S)>\delta} \ln \frac{P(S)-\delta}{Q(S)}$$

DEFINITION 11. We say that the family $\mathbb{M}$ of database access mechanisms satisfies ε-differential privacy under k-fold adaptive composition if for every adversary A, we have $D_\infty(V^0\|V^1) \le \varepsilon$ where V^b denotes the view of A in k-fold Composition Experiment b above.

(ε,δ)-differential privacy under k-fold adaptive composition instead requires that $D_\infty^\delta(V^0\|V^1) \le \varepsilon$.

C. Proof of Theorem 4

First, we describe the function $g_t(\cdot)$ in Algorithm 1 in a different way:

DEFINITION 12.

$$M_t = R(\theta_1,\dots,\theta_{d-1},\psi), \quad \text{where} \quad R(\theta) \text{ is a rotation matrix in } \mathbb{R}^d.$$

M_t satisfies a certain distribution such that:

$$M_t v \sim \begin{cases} \text{Uniform}(u:\langle u,v\rangle \ge 0) \ w.p. \frac{e^\epsilon}{e^\epsilon+1}, \\ \text{Uniform}(u:\langle u,v\rangle < 0) \ w.p. \frac{1}{e^\epsilon+1}, \end{cases}$$

where $v\in\mathbb{R}^d$.

Therefore, we have that $M_t v = g_t(v)$.

LEMMA 6. *In our algorithms, when $\eta \le 4\frac{e^\epsilon+1}{e^\epsilon-1}\mu$ we have*

$$\mathbb{E}_{g_t}\left\|\left(B_t - \eta\frac{n-1}{n} g_t(\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}))\right) - \left(B_t' - \eta\frac{n-1}{n} g_t(\nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k}))\right)\right\|_2 \le \|B_t - B_t'\|_2 .$$

where $\mathcal{D}^{-k},\mathcal{D}'^{-k}$ are datasets with $n-1$ data points by removing k-th data point from $\mathcal{D}$ and $\mathcal{D}'$.

Proof of Lemma 6 Due to the convexity and $\frac{1}{\mu}$-smoothness, we have the co-coercivity of gradient:

$$\langle \nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}) - \nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k}), B_t - B_t'\rangle \ge \mu\|\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}) - \nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k})\|_2^2,$$

$$\begin{aligned}
&\mathbb{E}_{g_t}\left\|\left(B_t - \eta\frac{n-1}{n} g_t(\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}))\right) - \left(B_t' - \eta\frac{n-1}{n} g_t(\nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k}))\right)\right\|_2^2 \\
&= \left\|(B_t - B_t') - \eta\frac{n-1}{2n}\frac{e^\epsilon-1}{e^\epsilon+1}\left(\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}) - \nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k})\right)\right\|_2^2 \\
&= \|B_t - B_t'\|_2^2 - \eta\frac{n-1}{n}\frac{e^\epsilon-1}{e^\epsilon+1}\langle B_t - B_t', \nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}) - \nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k})\rangle \\
&\quad + \left(\eta\frac{n-1}{2n}\frac{e^\epsilon-1}{e^\epsilon+1}\right)^2 \|\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}) - \nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k})\|_2^2 \\
&\le \|B_t - B_t'\|_2^2 + \left(\left(\eta\frac{n-1}{2n}\frac{e^\epsilon-1}{e^\epsilon+1}\right)^2 - \eta\mu\frac{n-1}{n}\frac{e^\epsilon-1}{e^\epsilon+1}\right)\|\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}) - \nabla\hat{\mathcal{L}}(B_t';\mathcal{D}'^{-k})\|_2^2 \\
&\le \|B_t - B_t'\|_2^2 \quad \text{when } \eta \le 4\frac{e^\epsilon+1}{e^\epsilon-1}\mu.
\end{aligned}$$

□

LEMMA 7 **(Bousquet and Elisseeff (2002))**. *For any symmetric learning algorithm $\mathcal{A}$, we have $\forall$ neighboring datasets $\mathcal{D}, \mathcal{D}'$:*

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n}[\mathcal{L}(\mathcal{A}(\mathcal{D})) - \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}))] = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n, d_i'}[\ell(\mathcal{A}(\mathcal{D}), d_i') - \ell(\mathcal{A}(\mathcal{D}'), d_i')], \tag{3}$$

where d_i' is the data point in $\mathcal{D}'$ which differs from $\mathcal{D}$.

LEMMA 8 **(Lemma 14.1 in Shalev-Shwartz and Ben-David (2014))**. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be an arbitrary sequence of vectors. Any algorithm with an initialization $\mathbf{w}^{(1)} = 0$ and an update rule of the form*

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$$

satisfies

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2.$$

With these necessary definitions and lemmas in place, we can now proceed to prove the upper bound of the algorithm.

Proof of Theorem 4 We analyze the algorithm's performance without projection, which gives a looser upper bound. We follow uniform stability and shrinking ERM framework to complete our proof. The regret can be decomposed into two parts,

$$\mathcal{R}(\mathbb{P}) = \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\mathcal{L}(B, \mathcal{D}) - \hat{\mathcal{L}}(B, \mathcal{D})\right] + \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\hat{\mathcal{L}}(B) - \hat{\mathcal{L}}(B^*)\right],$$

where the expectation $\mathbb{E}[\cdot]$ means taking expectation over algorithm's randomness, i.e., sampling noise vectors $\{g_t\}_{t=1}^T$, B^* is the optimal solution. We also omit $\mathbb{E}\left[\hat{\mathcal{L}}(B^*) - \mathcal{L}(B^*)\right]$since it has nothing to do with the algorithm.

We use B_t and B_t' to represent vectors in t-th iteration trained on neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, respectively. Their different data point is indexed by k, and let $\mathcal{D}^{-k}, \mathcal{D}'^{-k}$ denote datasets with $n-1$ data points by removing k-th data point. So $\mathcal{D}^{-k}$ and $\mathcal{D}'^{-k}$ are identical. We first conjecture that, with $\eta \leq 4\frac{e^\epsilon+1}{e^\epsilon-1}\mu$,

$$\mathbb{E}_{ED}\|B_t - B_t'\|_2 \leq \frac{2L\eta t}{n}, \quad \forall\, t = 1, \ldots, T.$$

We prove this conjecture by induction. When $t = 1$, by the setting of initial points, obviously it is true. Then suppose it is true for t-th iteration, it remains to check $(t+1)$-th iteration. $g_t(x) := \|x\|_2 G_t(x)$. By the update rule,

$$\|B_{t+1} - B_{t+1}'\|_2 = \left\| \left(B_t - \eta \|\nabla \hat{\mathcal{L}}_\mu(B_t; \mathcal{D})\| g_t(\nabla \hat{\mathcal{L}}_\mu(B_t; \mathcal{D}))\right) - \left(B_t' - \eta \|\nabla \hat{\mathcal{L}}_\mu(B_t'; \mathcal{D}')\| g_t(\nabla \hat{\mathcal{L}}_\mu(B_t'; \mathcal{D}))\right) \right\|_2$$

$$\leq \left\| \left(B_t - \eta \frac{n-1}{n} g_t(\nabla \hat{\mathcal{L}}_\mu(B_t; \mathcal{D}^{-k})) \right) - \left(B'_t - \eta \frac{n-1}{n} g_t(\nabla \hat{\mathcal{L}}_\mu(B'_t; \mathcal{D}'^{-k})) \right) \right\|_2 + \frac{\eta}{n} \left\| g_t(\nabla \ell_\mu(B_t; x_k, c_k)) - \nabla \ell_\mu(B'_t; x'_k, c'_k) \right\|_2 .$$

Then, the first term can be upper bounded by Lemma 6:

$$\mathbb{E}_{g_t} \left\| \left(B_t - \eta \frac{n-1}{n} g_t(\nabla \hat{\mathcal{L}}_\mu(B_t; \mathcal{D}^{-k})) \right) - \left(B'_t - \eta \frac{n-1}{n} g_t(\nabla \hat{\mathcal{L}}_\mu(B'_t; \mathcal{D}'^{-k})) \right) \right\|_2 \leq \left\| B_t - B'_t \right\|_2 ,$$

and the second term can be upper bounded by Lipschitz continuity arguments, i.e.,

$$\frac{\eta}{n} \left\| g_t(\nabla \ell_\mu(B_t; x_k, c_k) - \nabla \ell_\mu(B'_t; x'_k, c'_k)) \right\|_2 \leq \frac{\eta}{n} \left\| g_t \right\| \left\| \nabla \ell_\mu(B_t; x_k, c_k) - \nabla \ell_\mu(B'_t; x'_k, c'_k) \right\| \leq \frac{2L\eta}{n}.$$

Consequently,

$$\mathbb{E}_{g_t} \left\| B_{t+1} - B'_{t+1} \right\|_2 \leq \left\| B_t - B'_t \right\|_2 + \frac{2L\eta}{n} \leq \frac{2L\eta(t+1)}{n}.$$

Therefore, our conjecture is correct. With this conjecture, we are able to establish uniform stability:

$$\mathbb{E} \| \ell(B; x, c) - \ell(B'; x, c) \| \leq \mathbb{E} \left[L \cdot \frac{1}{T} \sum_{t=1}^{T} \left\| B_t - B'_t \right\|_2 + \mu D \right] \leq \frac{L^2 \eta (1+T)}{n} + \mu D, \quad \forall\ x, c.$$

where $D = \sup_{w \in S} \frac{1}{2} \|w\|_2 = \frac{1}{2} \mathbf{B}_w^2$.

Hence, by Lemma 7, we have

$$\mathbb{E} \left[\mathcal{L}(B) - \hat{\mathcal{L}}(B) \right] \leq \frac{L^2 \eta (1+T)}{n} + \mu D.$$

It remains to control the second term. Let us fix a dataset $\mathcal{D}$.

$$\begin{aligned}
\mathbb{E} \left[\hat{\mathcal{L}}(B) - \hat{\mathcal{L}}(B^*) \right] &\leq \mathbb{E} \left[\hat{\mathcal{L}}_\mu \left(\frac{1}{T} \sum_{t=1}^{T} B_t \right) - \hat{\mathcal{L}}_\mu(B^*) \right] + \mu D \\
&\overset{(a)}{\leq} \frac{1}{T} \cdot \mathbb{E} \left[\sum_{t=1}^{T} \left(\hat{\mathcal{L}}_\mu(B_t) - \hat{\mathcal{L}}_\mu(B^*) \right) \right] + \mu D \\
&\overset{(b)}{\leq} \frac{1}{T} \cdot \mathbb{E} \left[\sum_{t=1}^{T} \langle B_t - B^*, \nabla \hat{\mathcal{L}}_\mu(B_t) \rangle \right] + \mu D \\
&\overset{(c)}{=} \frac{1}{2T} \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \cdot \mathbb{E} \left[\sum_{t=1}^{T} \langle B_t - B^*, g_t(\nabla \hat{\mathcal{L}}_\mu(B_t)) \rangle \right] + \mu D \\
&\overset{(d)}{\leq} \frac{1}{2T} \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \cdot \mathbb{E} \left[\frac{\|B^*\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \left\| \nabla \hat{\mathcal{L}}_\mu \right\|_2^2 \right] + \mu D \\
&\overset{(e)}{\leq} \frac{1}{2} \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \left(\frac{\|B^*\|_2^2}{2T\eta} + \frac{\eta L^2}{2} \right) + \mu D.
\end{aligned}$$

(a), (b) follow directly from the convexity of $\hat{\mathcal{L}}$.

(c) follows the state of M_t in Definition 12, which is independent of B_t and the gradient $\nabla\hat{\mathcal{L}}_\mu$, i.e. $G_t = M_t\nabla\hat{\mathcal{L}}_\mu(B_t)$. Therefore, we can derive the following formula:

$$\begin{aligned}
\mathbb{E}\left[\langle B_t - B^*, g_t(\nabla\hat{\mathcal{L}}_\mu(B_t))\rangle\right] &= \mathbb{E}\left[\langle B_t - B^*, M_t\nabla\hat{\mathcal{L}}_\mu(B_t)\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\langle B_t - B^*, M_t\nabla\hat{\mathcal{L}}_\mu(B_t)\rangle \middle| B_t - B^*, \nabla\hat{\mathcal{L}}_\mu(B_t)\right]\right] \\
&= \mathbb{E}\left[(B_t - B^*)^T\mathbb{E}\left[M_t | B_t - B^*, \nabla\hat{\mathcal{L}}_\mu(B_t)\right]\nabla\hat{\mathcal{L}}_\mu(B_t)\right] \\
&= \mathbb{E}\left[(B_t - B^*)^T\mathbb{E}[M_t]\nabla\hat{\mathcal{L}}_\mu(B_t)\right] \\
&= \frac{1}{2}\frac{e^\varepsilon}{e^\varepsilon+1}\mathbb{E}\left[\langle B_t - B^*, \nabla\hat{\mathcal{L}}_\mu(B_t)\rangle\right] + \frac{1}{2}\frac{1}{e^\varepsilon+1}\mathbb{E}\left[\langle B_t - B^*, -\nabla\hat{\mathcal{L}}_\mu(B_t)\rangle\right] \\
&= \frac{1}{2}\frac{e^\varepsilon-1}{e^\varepsilon+1}\mathbb{E}\left[\langle B_t - B^*, \nabla\hat{\mathcal{L}}_\mu(B_t)\rangle\right].
\end{aligned}$$

(d) follows Lemma 8 directly.

So,

$$\begin{aligned}
\mathcal{R}(\mathbb{P}) =& \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\mathcal{L}(B,\mathcal{D}) - \hat{\mathcal{L}}(B,\mathcal{D})\right] + \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\hat{\mathcal{L}}(B) - \hat{\mathcal{L}}(B^*)\right] \\
\leq& \frac{L^2(1+T)\eta}{n} + \frac{e^\varepsilon-1}{e^\varepsilon+1}\left(\frac{\|B^*\|_2^2}{4T\eta} + \frac{\eta L^2}{4} + \frac{\eta}{2}D\right).
\end{aligned}$$

And furthermore, since we have already defined that $\|B\|_2 \leq \mathbf{B}_B,\ D = \frac{1}{2}\mathbf{B}_w^2$, we therefore choose the step size to be

$$\eta = \frac{\sqrt{\frac{e^\varepsilon-1}{e^\varepsilon+1}\frac{\mathbf{B}_B^2}{4T}}}{\sqrt{\frac{L^2(1+T)}{n} + \frac{e^\varepsilon-1}{e^\varepsilon+1}\left(\frac{L^2}{4} + \frac{\mathbf{B}_w^2}{4}\right)}}, \quad T = n^2$$

then

$$\begin{aligned}
\mathcal{R}(\mathbb{P}) &\leq 2\sqrt{\frac{e^\varepsilon-1}{e^\varepsilon+1}\frac{\mathbf{B}_B^2}{4n^2}\left(\frac{L^2(1+n^2)}{n} + \frac{e^\varepsilon-1}{e^\varepsilon+1}\left(\frac{L^2}{4} + \frac{\mathbf{B}_w^2}{4}\right)\right)} \\
&= \mathcal{O}\left(\sqrt{\frac{1}{n}} + \frac{e^\varepsilon-1}{e^\varepsilon+1}\frac{1}{n}\right)
\end{aligned}$$

□

D. Proof of Theorem 5

Proof of Theorem 5 First, define the domain of $\hat{B}$ as $\{\hat{B} \in \mathbb{R}^{d\times m} : \hat{B}_{ij} \geq 0\}$.

Consider $c^{(j)} = \hat{B}^{(j)}x^{(j)}$, where $\hat{B}^{(j)}$ has value 1 only at position $(j,1)$ and 0 elsewhere, and $x^{(j)} = (1,\ldots,1)^T$ for $1 \leq j \leq d$. For example: when $m = d = 2$, $\hat{B}^{(1)} = \begin{pmatrix}1&0\\0&0\end{pmatrix}$, $\hat{B}^{(2)} = \begin{pmatrix}0&0\\1&0\end{pmatrix}$. Thus, for any $i \neq j$, $||c^{(i)} - c^{(j)}||_1 = 2$.

[Part I] Transformation of $\mathcal{L}_{SPO+}$

$\mathcal{L}_{SPO+} = \mathbb{E}_{D\sim\mathbb{P}^n}[\mathcal{L}(\hat{B})]$. Assuming $\mathbb{P}^n$ is a uniform discrete distribution, i.e., $P((c,x)=(c_i,x_i)) = \frac{1}{n}$, for $1 \le i \le n$, we have:

$$\mathcal{L}_{SPO+} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{SPO+}(c_i,x_i)$$

which is the Empirical Risk Minimization (ERM). We consider the domain of the decision function w to be $S = \{w \ge -1\} = \{Iw \ge -\mathbf{1}\}$.

From Proposition 7 in Elmachtoub and Grigas (2022) we can transform the original problem into:

$$\min_{\hat{B},P}\mathcal{L}_{SPO+} = \min_{\hat{B},P}\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}^T(2\hat{B}x_i - c_i) + 2(w^*(c_i)x_i^T)\cdot\hat{B} - z^*(c_i)] \qquad (*)$$

In (*), consider $x_1 = \cdots = x_n = \mathbf{1}^T$ and $\{\hat{B}_i\}_{i=1}^n \subset \{\hat{B}^{(j)}\}_{j=1}^m$, $c_i = \hat{B}_i x_i$, $1 \le i \le n$. Taking $w^*(c_i) = -c_i$, it is easy to see that $\frac{1}{n}\sum_{i=1}^n \hat{B}_i$ is an optimal solution.

[Part II] Proof by contradiction (inspired by the proof of Lemma V.1 in Bassily et al. (2014))

Construst $D^{(j)}$ containing n copies of $\{\hat{B}^{(j)}x^{(j)}, x^{(j)}\}$, $1 \le j \le d$. Let $n^* = 100\frac{e^\varepsilon - 1}{e^\varepsilon + 1}$, let $\hat{c}_j = \hat{B}_j x^{(j)}$, $c_j^* = \hat{B}_j^* x^{(j)}$, where $\hat{B}_j$ is the result of the algorithm applied to $D^{(j)}$, $\hat{B}_j^*$ is the optimal solution based on $D^{(j)}$.

(1) When $n = n^*$, assume for all $1 \le j \le d$, $P(||\hat{c}_j - c_j^*||_1 < 1) \ge \frac{1}{2}$

i.e. $\mathcal{M}(c_j^*) := \{\hat{c}_j : ||\hat{c}_j - c_j^*||_1 < 1\}$, $P(\hat{c}_j \in \mathcal{M}(c_j^*)) \ge \frac{1}{2}$.

Since the algorithm satisfies ε-differential privacy, we can obtain:

$\frac{1}{2}e^{-\varepsilon n} \le P(\hat{c}_1 \in \mathcal{M}(c_j^*))$, where $\mathcal{M}(c_j^*)$ is disjoint with respect to j.

Summing over j, $d\cdot\frac{1}{2}e^{-\varepsilon n} \le \sum_{j=1}^n P(\hat{c}_1 \in \mathcal{M}(c_j^*)) \le 1$.

When p is large enough, n must also be very large. In this case, $n \le n^*$ leads to a contradiction.

Therefore, $\exists j$ s.t. $||\hat{c}_j - c_j^*||_1 \ge 1$ with probability at least $\frac{1}{2}$.

(2) When $n > n^*$,

Construct $\tilde{D}^{(j)}$ containing n copies of $(B^{(j)}x^{(j)}, x^{(j)})$ and $\lfloor\frac{n-n^*}{2}\rfloor$ copies of $(x^{(j)}, x^{(j)})$ and $\lfloor\frac{n-n^*}{2}\rfloor$ copies of $(-x^{(j)}, x^{(j)})$. Note that c_j^*, $\hat{c}_j$ are still obtained from $D^{(j)}$, $\tilde{D}^{(j)}$ constructed according to the algorithm in (1).

Define a new algorithm $\tilde{\mathcal{A}}$:

Input $D^{(j)}$. Add $\lfloor\frac{n-n^*}{2}\rfloor$ copies of $(x^{(j)}, x^{(j)})$ and $\lfloor\frac{n-n^*}{2}\rfloor$ copies of $(-x^{(j)}, x^{(j)})$ to $D^{(j)}$, yielding $\tilde{D}^{(j)}$.

Output $\frac{n}{n^*}(\mathcal{A}(\tilde{D}^{(j)}) - \alpha)$, where

$$\alpha = \begin{cases}\mathbf{I}, & n-n^* \text{ odd}\\ \mathbf{0}, & n-n^* \text{ even}\end{cases}$$

It's easy to find that $\tilde{\mathcal{A}}$ satisfies ε-differential privacy. Assume that for all $\forall\ 1 \le j \le d$, $P(||\mathcal{A}(\tilde{D}^{(j)})x^{(j)} - \tilde{c}_j^*||_1 < n^*) \ge \frac{1}{2}$.

$\Rightarrow P(||\tilde{\mathcal{A}}(D^{(j)})x^{(j)} - c_j^*||_1 < 1) \ge \frac{1}{2}$.

$\Rightarrow$ From (1) we can derive a contradiction. Therefore $||c_j - c_j^*||_1 \ge n^* = \Omega(\frac{e^\varepsilon - 1}{e^\varepsilon + 1})$ w.p. at least $\frac{1}{2}$. In conclusion, $\exists j$ s.t. $||\hat{c}_j - c_j^*||_1 \ge \frac{n^*}{n} = \Omega(\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\frac{1}{n})$.

[Part III] From solution to loss

$$\mathcal{L}_{\text{SPO+}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}^T(2\hat{B}\mathbf{x}_i - \mathbf{c}_i) + 2(\mathbf{w}^*(\mathbf{c}_i)\mathbf{x}_i^T)\cdot\hat{B} - z^*(\mathbf{c}_i)\right] \tag{4}$$

$$\begin{aligned}\mathcal{L}_{\text{SPO+}}(\hat{c}) - \mathcal{L}_{\text{SPO+}}(c^*) &= \frac{1}{n}\sum_{i=1}^{n}\left[2\mathbf{1}^T(\hat{B} - B^*)\mathbf{x}_i + 2(\mathbf{w}^*(\mathbf{c}_i)\mathbf{x}_i^T)\cdot(\hat{B} - B^*)\right] \\ &= \frac{2}{n}\sum_{i=1}^{n}\left[(\mathbf{1} + \mathbf{w}^*(\mathbf{c}_i))^T(\hat{C} - C^*)\right]\end{aligned}$$

$$\mathbb{E}_{\text{EG}}\left[\mathcal{L}_{\text{SPO+}}(\hat{c}) - \mathcal{L}_{\text{SPO+}}(c^*)\right] = \Omega\left(\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\cdot\frac{1}{n}\right) \tag{5}$$

In addition, we know from Bassily et al. (2019) that the lower bound of stochastic convex optimization is $\Omega(\frac{1}{\sqrt{n}})$, thus we have that

$$\mathbb{E}_{\text{EG}}\left[\mathcal{L}_{\text{SPO+}}(\hat{c}) - \mathcal{L}_{\text{SPO+}}(c^*)\right] = \Omega\left(\frac{1}{\sqrt{n}} + \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\cdot\frac{1}{n}\right)$$

□

E. Proof of Theorem 7

Proof of Theorem 7 Still, we consider the risk decomposing into two parts,

$$\mathcal{R}(\mathbb{P}) = \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\mathcal{L}(B,\mathcal{D}) - \hat{\mathcal{L}}(B,\mathcal{D})\right] + \mathbb{E}_{ED,\mathcal{D}\sim\mathbb{P}}\left[\hat{\mathcal{L}}(B) - \hat{\mathcal{L}}(B^*)\right],$$

For the first part, we still have $\left\|B_{t+1} - B'_{t+1}\right\|_2 \le \frac{2L\eta(t+1)}{n}$ because only the process of mathematical induction is related to the type of noise (Exponential or Gaussian), and this process can be rewritten as:

$$\begin{aligned}\|B_{t+1} - B'_{t+1}\|_2 &= \left\|\left(B_t - \eta(\nabla\hat{\mathcal{L}}(B_t;\mathcal{D})) + z_t\right) - \left(B'_t - \eta(\nabla\hat{\mathcal{L}}(B'_t;\mathcal{D})) + z_t\right)\right\|_2 \\ &\le \left\|\left(B_t - \eta\frac{n-1}{n}\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}))\right) - \left(B'_t - \eta\frac{n-1}{n}\nabla\hat{\mathcal{L}}(B'_t;\mathcal{D}^{-k}))\right)\right\|_2 \\ &\quad + \frac{\eta}{n}\left\|\nabla\ell(B_t;x_k,c_k) - \nabla\ell(B'_t;x'_k,c'_k)\right\|_2.\end{aligned}$$

where $z_t \sim \mathcal{N}(0,\sigma^2 I_d)$. Also, with Lemma 3.6 in Hardt et al. (2016) we know that

$$\left\|\left(B_t - \eta\frac{n-1}{n}\nabla\hat{\mathcal{L}}(B_t;\mathcal{D}^{-k}))\right) - \left(B'_t - \eta\frac{n-1}{n}\nabla\hat{\mathcal{L}}(B'_t;\mathcal{D}^{-k}))\right)\right\|_2 \le \|B_t - B'_t\|_2$$

And

$$\frac{\eta}{n}\left\|\nabla\ell(B_t;x_k,c_k)-\nabla\ell(B'_t;x'_k,c'_k)\right\|\le\frac{2L\eta}{n}.$$

Therefore we have $\left\|B_{t+1}-B'_{t+1}\right\|_2\le\frac{2L\eta(t+1)}{n}$, thus the first part is also bounded with $\frac{L^2\eta(1+T)}{n}+\mu D$.

For the second part, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathcal{L}}(B)-\hat{\mathcal{L}}(B^*)\right]&\le\frac{1}{T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\langle B_t-B^*,\nabla\hat{\mathcal{L}}_\mu(B_t)\rangle\right]+\mu D\\
&=\frac{1}{T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\langle B_t-B^*,\nabla\hat{\mathcal{L}}_\mu(B_t)+z_t\rangle\right]+\mu D\\
&\le\frac{1}{T}\cdot\mathbb{E}\left[\frac{\|B^*\|^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\left\|\nabla\hat{\mathcal{L}}_\mu(\hat{B}_t)\right\|^2+\|z_t\|^2\right]+\mu D,\\
&\le\frac{1}{T}\cdot\mathbb{E}\left[\frac{\|B^*\|^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\left\|\nabla\hat{\mathcal{L}}_\mu(\hat{B}_t)\right\|^2\right]+\frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|z_t\|_2^2\right]+\mu D,\\
&\le\frac{\|B^*\|^2}{2T\eta}+\frac{\eta L^2}{2}+\frac{\eta d\sigma^2}{2}+\frac{\eta}{2}D.
\end{aligned}$$

where $D=\frac{1}{2}\mathbf{B}_w^2$. Naturally, we take

$$\eta=\frac{\sqrt{\frac{\|\mathbf{B}_B\|^2}{2T}}}{\sqrt{\frac{L^2}{2}+\frac{d\sigma^2}{2}+\frac{1}{4}\mathbf{B}_w^2+\frac{L^2(1+T)}{n}}},\quad T=n^2$$

then

$$\mathcal{R}(\mathbb{P})=\mathcal{O}\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\ln(1/\delta)}}{n\epsilon}\right).$$

□

F. Proof of Theorem 8

The proof follows identically as the proof of Theorem 5 except by setting $n^*=100\frac{\sqrt{d\ln(1/\delta)}}{\epsilon}$.

□

G. Omitted Proofs in Section 6

Proof of Lemma 5 Since $c-2\mathbb{E}[c|x]\le 0$ a.s., we have

$$\begin{aligned}
\mathbb{E}_{c,x}\left[\max_{w\in S}\{c^Tw-2\mathbb{E}[c|x]^Tw\}\right]&=\mathbb{E}_{c,x}\left[(c-2\mathbb{E}[c|x])^T\min_{w\in S}w\right]\\
&=\mathbb{E}_x\left[\max_{w\in S}\{-\mathbb{E}[c|x]^Tw\}\right]
\end{aligned}$$

□

Proof of Theorem 9

$$\text{Left side} = \mathbb{E}_{c,x}\left[\max_{w\in S}\{c^T w - 2\mathbb{E}[c|x]^T w\}\right] + \mathbb{E}_{c,x}\left[2\mathbb{E}[c|x]^T w^*(c)\right] - \mathbb{E}_{c,x}\left[z^*(c)\right]$$

We begin by analyzing the first term:

$$\mathbb{E}_{c,x}\left[\max_{w\in S}\{c^T w - 2\mathbb{E}[c|x]^T w\}\right] = \mathbb{E}_x\left[\max_{w\in S}\left\{\mathbb{E}[c|x]^T w - 2\mathbb{E}[c|x]^T w\right\}\right] = \mathbb{E}_x\left[-\mathbb{E}[c|x]^T w^*(\mathbb{E}[c|x])\right]$$

The first equality holds because of Lemma 5. Next, we move on to the second term:

$$\mathbb{E}_{c,x}\left[2\mathbb{E}[c|x]^T w^*(c)\right] = \mathbb{E}_x\left[2\mathbb{E}[c|x]^T \mathbb{E}[w^*(c)|x]\right] \le \mathbb{E}_x\left[2\mathbb{E}[c|x]^T w^*(\mathbb{E}[c|x])\right],$$

the inequality holds because

$$\mathbb{E}[w^*(c)|x] = \int \left(\arg\min_{w\in S} c^T w\right) p(c|x)dc \le \arg\min_{w\in S}\left(\int c^T p(c|x)dc\right) w = w^*(\mathbb{E}[c|x])$$

This completes the evaluation of the left side. Now, turning to the right side:

$$\text{Right side} = \mathbb{E}_{c,x}\left[\max_{w\in W^*(\mathbb{E}[c|x])} c^T w\right] - \mathbb{E}\left[z^*(c)\right] = \mathbb{E}_{c,x}\left[c^T w^*(\mathbb{E}[c|x])\right] - \mathbb{E}[z^*(c)]$$

With both sides clearly defined, we observe that the left side is less than or equal to the right side. Moreover, from Lemma 4, we know that the left side is also greater than or equal to the right side. Therefore, we conclude that equality holds.

□

Proof of Corollary 1 $\ell_{SPO+}(\hat{B}) \ge \ell_{SPO}(\hat{B})$, $\mathbb{E}_{c,x}[\ell_{SPO+}\mathbb{E}[c|x]] = \mathbb{E}_{c,x}[\ell_{SPO}\mathbb{E}[c|x]]$. Therefore we have the inequality.

□