Differentially Private Synthetic Control

Saeyoung Rho
Columbia University

Rachel Cummings
Columbia University

Vishal Misra
Columbia University

Abstract

Synthetic control is a causal inference tool used to estimate the treatment effects of an intervention by creating synthetic counterfactual data. This approach combines measurements from other similar observations (i.e., *donor pool*) to predict a counterfactual time series of interest (i.e., *target unit*) by analyzing the relationship between the target and the donor pool before the intervention. As synthetic control tools are increasingly applied to sensitive or proprietary data, formal privacy protections are often required. In this work, we suggest the first algorithms for differentially private synthetic control with explicit error bounds based on the analysis of the sensitivity of the synthetic control query. Our approach builds upon tools from non-private synthetic control and differentially private empirical risk minimization. We empirically evaluate the performance of our algorithms and show favorable results in a variety of parameter regimes.

1 INTRODUCTION

The fundamental problem of causal inference is that for an individual unit, we can only observe one of the relevant outcomes – with a particular treatment or without (Rubin, 1974). To estimate the (causal) effect of a treatment, one has to produce a counterfactual of the test arm, which is typically done at a population- and distributional-level via randomized control trials (RCTs) and A/B testing, yielding average treatment effects. However, controlled trials are often impossible to implement, and only observational data are available. Synthetic control is a powerful causal inference tool to estimate the treatment effect of interventions using only observational data. It has been used both at an aggregate population level (e.g., countries/cities/cohorts of patients etc.), as well as at an individual unit level (Amjad et al., 2019; Agarwal et al., 2021a), and has been called "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens, 2017).

Recently, synthetic control has increasingly been used in clinical trials where running a randomized control trial presents logistical challenges (e.g., rare diseases), or ethical issues (e.g., oncology trials enrolling patients for placebos with life threatening diseases) (Thorlund et al., 2020). Synthetic control has been successfully used to achieve regulatory approval for new medical treatments for lung cancer (Petrone, 2018) and rare forms of leukemia (Gökbuget et al., 2016), where RCTs would otherwise have been impossible. Since these synthetic control analyses are deployed in real-world medical applications, preserving privacy of sensitive patient data is paramount.

Differential privacy by Dwork et al. (2006) has emerged as the de facto gold-standard in privacy-preserving data analysis. It is a mathematically rigorous parameterized privacy notion, which bounds the maximum amount that can be learned about any data donor based on analysis of her data. Differentially private algorithms have been designed for a wide variety of optimization, learning, and data-driven decision-making tasks, and have been deployed in practice by several major technology companies and government agencies. Despite the growing maturity of the differential privacy toolkit, the pressing need for a private synthetic control solution has thus far gone unaddressed.

In this paper, we propose the first algorithms for differentially private synthetic control (Algorithm 2 and 3). These two algorithms naturally extend existing non-private techniques for synthetic control by first privately estimating the regression coefficients $\hat{\boldsymbol{f}}$ that relate a (target) pre-intervention observation of interest $\boldsymbol{y_{pre}}$ to other similar (donor) observations X_{pre}. This is done using output perturbation and objective perturbation techniques for differentially private empirical risk minimization (DP-ERM) by Chaudhuri et al. (2011) and Kifer et al. (2012). The algorithm then combines the private regression coefficients $\hat{\boldsymbol{f}}$ with privatized post-intervention donor observations $\tilde{X}_{post}$ (also via output perturbation) to predict the post-intervention target outcome $\hat{\boldsymbol{y}}_{\boldsymbol{post}} = \tilde{X}_{post}^{\top}\hat{\boldsymbol{f}}$.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

We provide privacy and accuracy guarantees for each algorithm. For privacy (Theorems 3.1 and 3.4), although our algorithmic techniques rely on existing approaches, prior results on privacy do not apply in our setting. DP methods add noise that scales with the *sensitivity* of the function being computed, which is defined as the maximum change in the function's output that can be caused by changing a single donor's data. However, synthetic control performs a regression in a vertical way, treating each time point, rather than one donor's data point, as one sample – thus, the *transposed* setting changes the definition of neighboring databases, completely altering the impact of a single donor's data. The majority of our privacy analysis is devoted to computing sensitivity of this new method.

For accuracy guarantees (Theorems 3.2 and 3.5), we bound the root mean squared error (RMSE) of the algorithm's output compared to the post-intervention target signal. Our bounds are comparable to those for non-private synthetic control (e.g., Amjad et al. (2018)), and in Section C.1.1, we explicitly show that the RMSE of our algorithm relative to a non-private version is only greater by a factor of $O(1/\epsilon)$ for output perturbation, which is unavoidable in most analysis tasks. To better interpret our bounds in terms of natural problem parameters such as number of samples and length of observations, we also provide Corollaries C.5 and E.2, which give explicit closed-form upper bounds on the RMSE under mild assumptions on the underlying data distribution.

1.1 Related Work

Synthetic Control. Synthetic control (SC) was originally proposed byAbadie and Gardeazabal (2003) to evaluate the effects of intervention by creating synthetic counterfactual data. Its first application was measuring the economic impact of the 1960s terrorist conflict in Basque Country, Spain by combining GDP data from other Spanish regions prior to the conflict to construct a *synthetic* GDP dataset for Basque Country in the counterfactual world without the conflict. Synthetic control has since been applied to a wide array of topics such as estimating the effect of California's tobacco control program (Abadie et al., 2010), estimating the effect of the 1990 German reunification on per capita GDP in West Germany (Abadie et al., 2015), evaluating health policies (Kreif et al., 2016), forecasting weekly sales at Walmart stores (Amjad et al., 2019), and predicting cricket score trajectories (Amjad et al., 2019).

The core algorithm of synthetic control lies on finding a relationship between the target time series (e.g., GDP of Basque Country) and the donor pool (e.g., GDP of other Spanish regions). The original method by Abadie and Gardeazabal (2003) used linear regression with a simplex constraint on the weights: the regression coefficients should be non-negative and sum to one. Since its first introduction, the synthetic control literature has evolved to include a richer set of techniques, including tools to deal with multiple treated units (Abadie and L'Hour, 2021; Dube and Zipperer, 2015), to correct bias (Abadie and L'Hour, 2021; Ben-Michael et al., 2021), to use Lasso and Ridge regression instead of linear regression with simplex constraints (Doudchenko and Imbens, 2016; Amjad et al., 2018), and to incorporate matrix completion techniques (Athey et al., 2021; Amjad et al., 2018, 2019). See a review paper by Abadie (2021) for a detailed survey of these techniques.

The most relevant extension for our work is *robust synthetic control* (RSC) by Amjad et al. (2018), which comprises of two steps: first de-noising the data via hard singular value thresholding (HSVT), and then learning and projecting via regression. It assumes a latent variable model and applies HSVT before running the regression, which reduces the rank of the data. RSC also relaxes the simplex constraints on the regression coefficients and applies unconstrained Ridge regression. Because of the de-noising step, RSC can be viewed as an instantiation of principal component regression (PCR) and the possibility of differentially private PCR has been briefly discussed by Agarwal et al. (2021b). However, no formal algorithm or analysis has been put forth until this paper.

Differentially Private Empirical Risk Minimization. Chaudhuri et al. (2011) first proposed methods for differentially private empirical risk minimization (ERM) for supervised regression and classification. Our first algorithm uses the *output perturbation* method by Chaudhuri et al. (2011), which first computes coefficients to minimize the loss function between data features and labels, and then perturbs the coefficients using a high-dimensional variant of the Laplace Mechanism by Dwork et al. (2006). Our second algorithm uses the *objective perturbation* method by Chaudhuri et al. (2011) and Kifer et al. (2012), which adds noise directly to the loss function and then exactly optimizes the noisy loss. This method tends to provide better theoretical accuracy guarantees but requires the loss function to satisfy additional structural properties. These methods were later extended by Bassily et al. (2014) to include *gradient perturbation* in stochastic gradient descent, which uses a noisy version of randomly sampled points' contribution to the gradient at each update. This technique provides tighter error bounds, assuming Lipschitz convex loss and bounded optimization domain. Wang et al. (2017) followed up with a faster gradient perturbation algorithm that provided a tighter upper bound on error and lower gradient complexity.

Although the framework by Chaudhuri et al. (2011) is more general, the analysis and applications focused only on methods for binary classification. The analysis was later extended to include ridge regression by Cummings et al. (2015), which we use in our algorithms. Our algorithms

for differentially private synthetic control apply DP-ERM methods to a ridge regression loss function. However, synthetic control applies regression in the transposed dimension of the data (i.e., along columns rather than rows of the database), while privacy protections are still required along the rows, which requires novel analysis to ensure differential privacy and accuracy.

2 MODEL AND PRELIMINARIES

In this section, we first present our model (Section 2.1) and then provide relevant background on synthetic control (Section 2.2) and differential privacy (Section 2.3).

2.1 Our Model

Our model follows the synthetic control framework illustrated in Figure 1. We consider a database $X \in \mathbb{R}^{n \times T}$, also called the *donor pool*. The donor pool X consists of n time series, each observed at times $t = 1, \ldots, T$. We denote the column vectors of X as $\boldsymbol{x_1}, \cdots \boldsymbol{x_T} \in \mathbb{R}^n$, where each $\boldsymbol{x_t}$ contains observations from all donor time series at time t. We assume an intervention occurred at a known time $T_0 + 1 < T$. The first T_0 columns of X are collectively referred to as X_{pre}, and the remaining $T - T_0$ columns from data after the intervention are collectively denoted X_{post}, respectively corresponding to the pre- and post-intervention donor data. We are also given a *target unit* $\boldsymbol{y} \in \mathbb{R}^T$, which can be divided as $\boldsymbol{y_{pre}} = (y_1, \ldots, y_{T_0})$ and $\boldsymbol{y_{post}} = (y_{T_0+1}, \ldots, y_T)$.

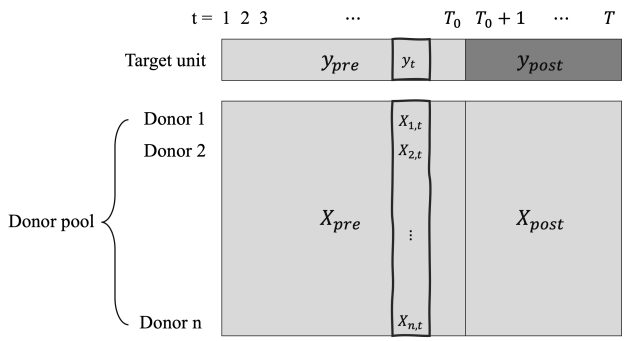

Figure 1: General data structure for synthetic control. The donor pool (X) and the target unit ($\boldsymbol{y}$) are divided into pre- and post-intervention periods. Synthetic control first performs a *vertical* regression using the pre-intervention column vectors $\boldsymbol{x_t}$ as features for the label y_t for $t \in [T_0]$ to estimate regression coefficients $\hat{\boldsymbol{f}}$, and then uses this to project the post-intervention column vectors and predict $\hat{\boldsymbol{y}}_{\boldsymbol{post}}$.

The underlying assumption is that time series in the donor pool that behave similarly to $\boldsymbol{y}$ before the intervention will remain similar after the intervention. In this paper, we use the latent variable model, the same as Amjad et al. (2018), for the underlying distribution of the data. Our donor data and target data are noisy versions of the true signal (denoted M and $\boldsymbol{m}$ respectively), and can be written as follows:

$$X = M + Z, \qquad \boldsymbol{y} = \boldsymbol{m} + \boldsymbol{z}, \tag{1}$$

where $Z \in \mathbb{R}^{n \times T}$ is a noise matrix where each element is sampled i.i.d. from some distribution with zero-mean, σ^2-variance, and support $[-s, s]$, and $\boldsymbol{z} \in \mathbb{R}^T$ is a noise vector with elements sampled from the same distribution.

The signals M and $\boldsymbol{m}$ can be expressed in terms of a latent function g:

$$M_{i,t} = g(\theta_i, \rho_t) \;\; m_t = g(\theta_0, \rho_t) \;, \forall i \in [n], t \in [T],$$

where θ_i and ρ_t are latent feature vectors capturing unit i's and time t's intrinsic characteristics, respectively. We note that if the intervention is effective, one would expect to see a change in ρ_t before and after T_0. We make no assumptions on the latent function g, except in Sections C.2 and E.2, where we assume M is low rank, i.e., $rank(M) = k$ for some $k \ll \min\{n, T\}$.

Finally, we assume a linear relationship between the features of $M_{i,t}$ and the label m_t at all times $t \in [T_0]$; that is, there exists an $\boldsymbol{f} \in \mathbb{R}^n$ such that,

$$m_t = \sum_{i=1}^{n} M_{i,t} f_i, \quad \text{for all } t \in [T_0]. \tag{2}$$

We assume that all entries of X, M, $\boldsymbol{y}$, and $\boldsymbol{m}$ lie in a bounded range, which we rescale to $[-1, 1]$ WLOG, and that $\boldsymbol{f}$ has ℓ_1-norm bounded by 1, as is standard in the synthetic control literature (Abadie and Gardeazabal, 2003). Formally, we assume:

$$|x_{i,t}| \leq 1, |M_{i,t}| \leq 1 \; \forall t \in [T], \; i \in [n],$$
$$|y_t| \leq 1, |m_t| \leq 1 \; \forall t \in [T], \, and \; ||\boldsymbol{f}||_1 = \sum_{k=1}^{n} |f_k| \leq 1. \tag{3}$$

2.2 Synthetic Control

The goal of synthetic control is to predict $\boldsymbol{y_{post}}$ given X and $\boldsymbol{y_{pre}}$. The general approach, outlined in algorithm 1, is to first use the pre-intervention data $\mathcal{D}_1 := (X_{pre}, \boldsymbol{y_{pre}})$ to learn an estimate $\hat{\boldsymbol{f}}$ of the true coefficient vector $\boldsymbol{f}$. For each $t \in [T_0]$, the column vector $\boldsymbol{x_t} = (X_{1,t}, \cdots, X_{n,t})^\top$ is treated as a feature vector for label y_t. This setup distinguishes synthetic control from the classic regression setting, as the regression is performed vertically rather than horizontally. The estimate $\hat{\boldsymbol{f}}$ is then used along with the post-intervention donor data to predict the counterfactual outcome of the target: $\hat{\boldsymbol{y}}_{\boldsymbol{post}} = X_{post}^\top \hat{\boldsymbol{f}}$, where $\hat{y}_t = \boldsymbol{x_t}^\top \hat{\boldsymbol{f}}$ $\forall t \in \{T_0 + 1, \cdots, T\}$.

Algorithm 1 Synthetic control framework $(X, \boldsymbol{y_{pre}}, T_0, J)$

Divide X into pre- and post-intervention observations

$$X = (X_{pre}, X_{post}) = \left(\begin{pmatrix} x_{1,1} & \cdots & x_{1,T_0} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,T_0} \end{pmatrix} \begin{pmatrix} x_{1,T_0+1} & \cdots & x_{1,T} \\ \vdots & \ddots & \vdots \\ x_{n,T_0+1} & \cdots & x_{n,T} \end{pmatrix} \right)$$

Step 1: Learn regression coefficients

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f} \in \mathbb{R}^n}{\arg\min}\, J(\boldsymbol{f}; X_{pre}, \boldsymbol{y_{pre}})$$

Step 2: Predict $\boldsymbol{y_{post}}$ via projection

Output $\hat{\boldsymbol{y}}_{\boldsymbol{post}} = X_{post}^\top \hat{\boldsymbol{f}} \in \mathbb{R}^{T-T_0}$

When synthetic control is used for evaluating treatment effects, it is assumed that the target received a different treatment from the donor pool, and the goal is to predict the counterfactual outcome under the alternative treatment. In this case, the treatment effect is evaluated as the difference between the observed $\boldsymbol{y_{post}}$ and the counterfactual prediction $\hat{\boldsymbol{y}}_{\boldsymbol{post}}$. Even if no intervention occurred at time $T_0 + 1$ (or if the target received the same treatment as the donor pool), then synthetic control can also be used to predict future observations of the target time series. In that case, accuracy can be measured as the difference between the actual observation $\boldsymbol{y_{post}}$ and the predicted $\hat{\boldsymbol{y}}_{\boldsymbol{post}}$. While the former task is the more common use-case for synthetic control, we focus our attentions in this work on the latter, in order to cleanly evaluate accuracy of our algorithms' predictions without confounding treatment effects.

The original synthetic control work by Abadie and Gardeazabal (2003) learned regression coefficients using ordinary linear regression with a simplex constraint on $\hat{\boldsymbol{f}}$, i.e., $\widehat{f}_i \geq 0 \ \forall i \in [n]$ and $\sum_{i \in [n]} \widehat{f}_i = 1$. Later works by Amjad et al. (2018, 2019); Doudchenko and Imbens (2016); Ben-Michael et al. (2021) used penalties such as Lasso, Ridge, and elastic net regularizers.

In this work, we use *Ridge regression*—with empirical loss $\mathcal{L}(\boldsymbol{f}; X, y) = \frac{1}{T_0}||y - X^\top \boldsymbol{f}||_2^2$ and an ℓ_2 regularizer $r(\boldsymbol{f}) = \frac{\lambda}{2T_0}||\boldsymbol{f}||_2^2$—to estimate $\hat{\boldsymbol{f}}$, which corresponds to the following regularized quadratic loss function:

$$J(\boldsymbol{f}; \mathcal{D}) = \mathcal{L}(\boldsymbol{f}; X_{pre}, \boldsymbol{y_{pre}}) + r(\boldsymbol{f}) \tag{4}$$

2.3 Differential Privacy

Differential privacy by Dwork et al. (2006) ensures that changing a single user's data will have only a bounded effect on the outcome of an algorithm. Specifically, it ensures that the distribution of an algorithm's output will be similar under two *neighboring databases* that differ only in a single data record. In the synthetic control setting, where the analysis goal is to predict the post-intervention target unit $\boldsymbol{y_{post}}$ using the donor pool X and its relationship to $\boldsymbol{y_{pre}}$, we aim to predict privacy of data records in X but not $\boldsymbol{y_{pre}}$, since the target will know their own pre-intervention data. Note that this is similar to the notion of *joint differential privacy* by Kearns et al. (2014), where personalized outputs to each user need not be private with respect to their own data, only to the data of others. Thus databases $D = (X, \boldsymbol{y})$ and $D' = (X', \boldsymbol{y})$ are considered neighboring in our setting if X and X' differ in at most one row and have the same target unit $\boldsymbol{y}$.

Definition 2.1 (Differential privacy (Dwork et al., 2006)) *A randomized algorithm $\mathcal{M}$ with domain $\mathcal{D}$ is (ϵ, δ)-differentially private for all $\mathcal{S} \subseteq$ Range$(\mathcal{M})$ and for all pairs of neighboring databases $D, D' \in \mathcal{D}$,*

$$\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in \mathcal{S}] + \delta,$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say $\mathcal{M}$ is ϵ-differentially private.

Definition 2.2 (ℓ_2 sensitivity) *The ℓ_2 sensitivity of a vector-valued function $\boldsymbol{f}$, denoted $\Delta \boldsymbol{f}$, is the maximum ℓ_2-norm change in the function's value between neighboring databases:*

$$\Delta \boldsymbol{f} = \max_{D, D' \text{ neighbors}} ||\boldsymbol{f}(D) - \boldsymbol{f}(D')||_2.$$

A common method for achieving ϵ-differential privacy for vector-valued functions is the *high-dimensional Laplace Mechanism* by Chaudhuri et al. (2011), which privately evaluates a function $\boldsymbol{f}$ on a dataset D by first evaluating $\boldsymbol{f}(D)$ and then adding a Laplace noise vector $\boldsymbol{v}$ sampled according to density $p(\boldsymbol{v}; a) \propto \exp\left(-\frac{||\boldsymbol{v}||_2}{a}\right)$, with parameter $a = \frac{\Delta \boldsymbol{f}}{\epsilon}$. Note that this is an extension of the (single-dimensional) Laplace Mechanism by Dwork et al. (2006), which would add Laplace noise with parameter $\Delta f / \epsilon$ to achieve ϵ-DP for real-valued queries. Alternatively, one can add Gaussian noise of mean 0 and standard deviation at least $\sqrt{2 \ln(1.25/\delta)} \Delta f / \epsilon$ to achieve (ϵ, δ)-DP.

Differential privacy is robust to *post-processing*, meaning that any downstream computation performed on the output of a differentially private algorithm will retain the same privacy guarantee. DP also *composes*, meaning that if an (ϵ_1, δ_1)-DP mechanism and an (ϵ_2, δ_2)-DP mechanism are performed on the same database, then the entire process is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.

3 DIFFERENTIALLY PRIVATE SYNTHETIC CONTROL (DPSC)

In this section, we present two algorithms for differentially private synthetic control, $DPSC_{out}$ (Algorithm 2)

and $DPSC_{obj}$ (Algorithm 3). Similar to non-private synthetic control algorithms (e.g., Algorithm 1), both algorithms are divided into two high-level steps: first the algorithm learns an estimate of the regression coefficients $\boldsymbol{f}$, and then it uses these coefficients to predict the post-intervention target unit $\boldsymbol{y_{post}}$. To ensure differential privacy of the overall algorithm, both of these steps must be performed privately. The second step remains the same for both, and only the first part differs: $DPSC_{out}$ adds privacy noise directly to the output of the algorithm (output perturbation), whereas $DPSC_{obj}$ perturbs the objective function and minimize the noisy objective (objective perturbation).

3.1 DPSC via Output Perturbation $DPSC_{out}$

Our first algorithm is $DPSC_{out}$ (Algorithm 2), which utilizes *output perturbation* to achieve differential privacy. The learning step of this algorithm formalizes synthetic control as an instance of empirical risk minimization with the Ridge regression loss function given in Equation (4). This enables us to apply the Output Perturbation method by Chaudhuri et al. (2011) for DP-ERM. The algorithm first learns non-private regression coefficients $\boldsymbol{f}^{reg}$ as in Algorithm 1. It then samples a noise vector $\boldsymbol{v}$ from a high-dimensional Laplace distribution with parameter $\Delta \boldsymbol{f}^{reg}/\epsilon_1$, as described in Section 2.3. Finally, the privatized regression coefficient vector is $\boldsymbol{f}^{out} = \boldsymbol{f}^{reg} + \boldsymbol{v}$.

The prediction step uses this coefficient vector to predict $\boldsymbol{y_{post}}$. A simple approach would be to directly predict $\hat{\boldsymbol{y}}_{\boldsymbol{post}} = X_{post}^{\top} \boldsymbol{f}^{out}$; however, this approach would not provide privacy for the post-intervention donor data X_{post}. Instead, we again apply the high-dimensional Laplace Mechanism to privatize X_{post} by adding a noise matrix W sampled from a high-dimensional Laplace distribution with parameter $\Delta X_{post}/\epsilon_2$. The privatized version of donor data is $\tilde{X}_{post} = X_{post} + W$, which is then used along with $\boldsymbol{f}^{out}$ to produce the private prediction of the post-intervention target unit: $\boldsymbol{y}^{out} = \tilde{X}_{post}^{\top} \boldsymbol{f}^{out}$.

The entire algorithm is then $(\epsilon_1 + \epsilon_2, 0)$-differentially private by composition of these two steps. We remark that the algorithm does not output $\boldsymbol{f}^{out}$, simply because this vector is typically not of interest in most cases, and is instead considered only an intermediate analysis step. However, this vector could be output if desired with no additional privacy loss because Step 1 of the $DPSC_{out}$ algorithm is ϵ_1-differentially private (Theorem B.1), and this privacy loss is already accounted for in the composition step.

We provide two main results on the privacy and accuracy of $DPSC_{out}$. First, Theorem 3.1 shows that this algorithm is differentially private. Although our algorithm relies on algorithmic techniques by Chaudhuri et al. (2011) for DP-ERM, the vertical regression setup in synthetic control requires novel sensitivity analysis for $\boldsymbol{f}^{reg}$, which constitutes the bulk of the work required to prove Theorem 3.1. The-

Algorithm 2 DPSC via Output Perturbation $DPSC_{out}(X_{pre}, X_{post}, \boldsymbol{y_{pre}}, n, T, T_0, \lambda, \epsilon_1, \epsilon_2)$

Step 1: Learn regression coefficients

Learn the regression coefficient $\boldsymbol{f}^{reg}$ using Ridge regression with parameter $\lambda \geq 0$:

$$\boldsymbol{f}^{reg} = \underset{\boldsymbol{f} \in \mathbb{R}^n}{\arg\min} \frac{1}{T_0} ||\boldsymbol{y_{pre}} - X_{pre}^{\top} \boldsymbol{f}||_2^2 + \frac{\lambda}{2T_0} ||f||_2^2.$$

Let $a = \frac{\Delta \boldsymbol{f}^{reg}}{\epsilon_1} = \frac{4T_0\sqrt{8+n}}{\lambda \epsilon_1}$

Sample $\boldsymbol{v}$ according to pdf $p(\boldsymbol{v}; a) \propto \exp\left(-\frac{||\boldsymbol{v}||_2}{a}\right)$

Let $\boldsymbol{f}^{out} = \boldsymbol{f}^{reg} + \boldsymbol{v}$

Step 2: Predict $\boldsymbol{y_{post}}$ via projection

Let $b = \frac{2\sqrt{T-T_0}}{\epsilon_2}$

Sample each entry of $W \in \mathbb{R}^{n \times (T-T_0)}$ i.i.d. according to pdf $p(W; b) \propto \exp\left(-\frac{||W||_F}{b}\right)$

Let $\tilde{X}_{post} = X_{post} + W$

Output $\boldsymbol{y}^{out} = \tilde{X}_{post}^{\top} \boldsymbol{f}^{out}$.

orem 3.2 shows that our $DPSC_{out}$ algorithm produces an accurate prediction of the post-intervention target unit, as measured by the standard metric of root mean squared error (RMSE) with respect to the true signal vector $\boldsymbol{m}$. In Section C.2, we also extend Theorem 3.2 to to remove the dependence on distributional parameters and provide an expression of RMSE that depends only on the input parameters, under some mild additional assumptions on the distribution of data. Full proofs for Theorems 3.1 and 3.2 are respectively presented in Sections B and C.

Theorem 3.1 *$DPSC_{out}$ of Algorithm 2 is $(\epsilon_1 + \epsilon_2, 0)$-differentially private.*

Theorem 3.2 *The estimator $\boldsymbol{y}^{out}$ output by Algorithm 2 satisfies:*

$$\begin{aligned} &RMSE(\boldsymbol{y}^{out}) \\ &\leq \frac{||M_{post}||_2}{\sqrt{T-T_0}} \left(\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] + \frac{4T_0\sqrt{8+n}}{\lambda \epsilon_1} \right) \\ &+ \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2} \right) \left(\sqrt{n}\psi + \frac{4T_0\sqrt{8+n}}{\lambda \epsilon_1} \right), \end{aligned}$$

where $||\boldsymbol{f}^{reg}||_\infty \leq \psi$ for some $\psi > 0$, and RMSE is the root mean squared error of the estimator, defined as $RMSE(\boldsymbol{y}^{out}) = \frac{1}{\sqrt{T-T_0}} \mathbb{E}[||\boldsymbol{y}^{out} - \boldsymbol{m}_{post}||_2]$.

Remark 3.3 *The accuracy bound grows as $O(n)$, which is shown to be necessary in Section B.1.1. While this might be undesirable in most other learning domains, n does not grow with the problem size in synthetic control settings for several reasons. Typically, M is assumed to be a low-rank*

matrix and hence X is approximately *low rank (Amjad et al., 2018, 2019). This is not only an assumption, but true in most cases (Udell and Townsend, 2019). Therefore, there exists a saturation point where adding additional donors does not meaningfully improve accuracy (see Section 3.3 for more details). The remaining dependence on T_0 can be handled by setting $\lambda = O(T_0)$ (see Section C.1.1).*

3.2 DPSC via Objective Perturbation $DPSC_{obj}$

We next present our second algorithm for differentially private synthetic control, $DPSC_{obj}$ (Algorithm 3), based on objective perturbation. While Step 2 remains unchanged relative to Algorithm 2, Step 1 is modified to perturb the objective function itself and then exactly optimize the perturbed objective, instead of first computing the optimal non-private coefficients and then adding noise. Objective perturbation has been shown to outperform output perturbation in the standard private ERM setting when the loss function is strongly convex (Chaudhuri et al., 2011).

The algorithm augments the objective function with two terms. The first is an additional regularization term to ensure $\frac{\lambda+\Delta}{T_0}$-strong convexity (compared to $\frac{\lambda}{T_0}$ with the regularization term of Algorithm 2). The Δ parameter is tuned by the algorithm to ensure that it can still satisfy (ϵ_1, δ)-DP in Step 1, even when ϵ_1 is small. The second is the noise term $\boldsymbol{b}^\top \boldsymbol{f}$ to ensure privacy, where $\boldsymbol{b}$ is sampled from a high-dimensional Laplace distribution if $(\epsilon, 0)$-DP is desired (i.e., if $\delta = 0$), and from a multi-variate Gaussian distribution if (ϵ, δ)-DP is desired (i.e., if $\delta > 0$).

The algorithm then exactly optimizes this new objective function, where the noise term $\boldsymbol{b}$ ensures that this minimization satisfies differential privacy. Although the algorithmic procedure in Step 1 is similar to that of Objective Perturbation algorithms for DP-ERM by Chaudhuri et al. (2011) and Kifer et al. (2012), the sensitivity and privacy analysis again requires substantial novelty because the definition of neighboring databases change and previous work cannot be immediately applicable to the transposed regression setting. Finally, Algorithm 3 maintains the same Step 2 process as Algorithm 2 to predict $\boldsymbol{y_{post}}$, based on $\boldsymbol{f}^{obj}$ computed from Step 1. Algorithm 3 is $(\epsilon_1 + \epsilon_2, \delta)$-DP by composition of privacy guarantees from these two steps.

$DPSC_{obj}$ requires an additional parameter c that is used in the analysis to bound the maximum absolute eigenvalue of $2(X'_{pre}X'^{\top}_{pre} - X_{pre}X^{\top}_{pre})$, which is closely related to $||\nabla \mathcal{L}(\boldsymbol{f})||_2$. Because X_{pre} and X'_{pre} are neighboring databases, the matrix of interest will only have one column and one row that are non-zero. In our setting, we use the fact that all entries of X are in $[-1, 1]$ to derive an upper bound on this matrix and its eigenvalues. In general, an analyst can use domain expertise or prior knowledge of the data distribution to choose an appropriate value of c.

Algorithm 3 DPSC via Objective Perturbation $DPSC_{obj}(X_{pre}, X_{post}, y_{pre}, n, T, T_0, \lambda, \epsilon_1, \epsilon_2, \delta, c)$

Step 1: Learn regression coefficients

if $\epsilon_1 > \log(1 + \frac{2c}{\lambda} + \frac{c^2}{\lambda^2})$ **then**

  Let $\epsilon_0 = \epsilon_1 - \log(1 + \frac{2c}{\lambda} + \frac{c^2}{\lambda^2})$ and $\Delta = 0$

else

  $\epsilon_0 = \frac{\epsilon_1}{2}$ and $\Delta = \frac{c}{e^{(\epsilon_1/4)}-1} - \lambda$

end if

if $\delta > 0$ **then**

  Sample $\boldsymbol{b} \in \mathbb{R}^n$ from $\mathcal{N}(0, \beta^2 I_n)$,

  where $\beta = \frac{4T_0\sqrt{8+n}\sqrt{2\log\frac{2}{\delta}+2\epsilon_0}}{\epsilon_0}$

else

  Sample $\boldsymbol{b} \in \mathbb{R}^n$ from $p(\boldsymbol{b}; \beta) \propto \exp\left(-\frac{||\boldsymbol{b}||_2}{\beta}\right)$,

  where $\beta = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$

end if

Learn $\boldsymbol{f}^{obj}$ by minimizing

$$J = \frac{1}{T_0}||y_{pre} - X^{\top}_{pre}\boldsymbol{f}||_2^2 + \frac{\lambda+\Delta}{2T_0}||\boldsymbol{f}||_2^2 + \frac{1}{T_0}\boldsymbol{b}^\top \boldsymbol{f}.$$

Step 2: Predict $\boldsymbol{y_{post}}$ via projection

Let $b = \frac{2\sqrt{T-T_0}}{\epsilon_2}$

Sample each entry of $W \in \mathbb{R}^{n\times(T-T_0)}$ i.i.d. according to pdf $p(W; b) \propto \exp\left(-\frac{||W||_F}{b}\right)$

Let $\tilde{X}_{post} = X_{post} + W$

Output $\boldsymbol{y}^{obj} = \tilde{X}^{\top}_{post}\boldsymbol{f}^{obj}$

We provide two main results on the privacy and accuracy of $DPSC_{obj}$. First, Theorem 3.4 shows that our algorithm is differentially private. To prove privacy in Step 1, we must consider two cases based on the value of Δ, which adds additional strong convexity to the loss function if it is needed. The privacy budget must be allocated differently within the analysis in the two cases of $\Delta = 0$ and $\Delta > 0$.

Theorem 3.5 shows that $DPSC_{obj}$ produces an accurate prediction of the post-intervention target unit, as measured as RMSE between its output $\boldsymbol{y}^{obj}$ and the target unit's post-intervention signal vector $\boldsymbol{m_{post}}$. As with $DPSC_{out}$, we also extend Theorem 3.5 in Section E.2 to provide an explicit closed-form bound on RSME that does not depend on the distributional parameters. Full proofs for for Theorems 3.4 and 3.5, along with their extensions, are respectively presented in Sections D and E.

Theorem 3.4 *$DPSC_{obj}$ of Algorithm 3 is $(\epsilon_1 + \epsilon_2, \delta)$-differentially private.*

Theorem 3.5 *The estimator $\boldsymbol{y}^{obj}$ output by Algorithm 3*

satisfies:

$$RMSE(\boldsymbol{y}^{obj}) \leq \frac{||M_{post}||_2}{\sqrt{T-T_0}}\Bigg(\mathbb{E}[||(\boldsymbol{f}^{reg}-\boldsymbol{f})||_2] + \frac{2}{\lambda+\Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta\neq 0}\left(\frac{1}{\lambda}+\frac{1}{\lambda+\Delta}\right)2T_0^2\sqrt{n}\Bigg) + \left(\sqrt{n\sigma^2}+\frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \frac{2}{\lambda+\Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta\neq 0}\left(\frac{1}{\lambda}+\frac{1}{\lambda+\Delta}\right)2T_0^2\sqrt{n}\right),$$

where $||\boldsymbol{f}^{reg}||_\infty \leq \psi$ *for some* $\psi > 0$*, and* $\mathbb{E}[||\boldsymbol{b}||_2] = \sqrt{\frac{nT_0 4\sqrt{8+n}\sqrt{2\log\frac{2}{\delta}+\epsilon_0}}{\epsilon_0}}$ *for Gaussian noise (*$\delta > 0$ *case) and* $\mathbb{E}[||\boldsymbol{b}||_2] = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$ *for Laplace noise (*$\delta = 0$ *case), and* ϵ_0*, and* Δ *are computed internally by the algorithm.*

As in Section 3.1, we remark that while the accuracy bound of Theorem 3.5 grows as $O(n)$, in our setting n does not typically grow substantially with the problem size, both in theory (Amjad et al., 2018, 2019) and in practice (Udell and Townsend, 2019).

3.3 Comparing DP-ERM and DPSC

In this section, we compare the results of DP-ERM by Chaudhuri et al. (2011) with our approach. Consider a Ridge regression task in p-dimensional space with q samples (i.e., covariates $x_k \in \mathbb{R}^p$ and labels $y_k \in \mathbb{R}$, $\forall k \in [q]$). The regression coefficient $\theta \in \mathbb{R}^p$ is learned by a standard empirical risk minimization process with a regularizer $\propto \lambda||\theta||_2^2$. In a typical regression setup where the privacy goal is to protect one sample x_k, corresponding to one individual's data, the sensitivity of the coefficient is $\Delta\theta = \frac{2}{q\lambda}$ (Chaudhuri et al., 2011). It does not depend on the dimension p, and the sensitivity decreases as the number of samples q increases. Intuitively, adding or removing one person's data should exhibit diminishing marginal effect on the final model θ as the training sample size grows.

On the other hand, in our transposed setting of synthetic control, the privacy goal is to protect the i-th entry of each x_k (i.e., an individual's data are spread across all samples), the sensitivity is $\Delta\theta = \frac{4q\sqrt{8+p}}{\lambda}$ (Lemma 3.7). In this setting, each dimension of the coefficient θ captures how important the corresponding donor is for explaining the target; hence the impact of changing i-th person's data will have a significant on the i-th dimension of θ, regardless of the number of individuals in the donor pool. This is at the crux of why it is more difficult to guarantee privacy in the transposed setting of synthetic control, relative to the standard regression setting.

3.4 Proof Sketch for Privacy Guarantees

In this section, we outline the proof for privacy guarantees for both algorithms. The full versions with all omitted proofs are presented in Appendices B and D.

The proof of Theorem 3.1 relies on the privacy of $\boldsymbol{f}^{out}$ in the learning phase, and then $\tilde{X}_{post}$ in the prediction phase. At a high level, $\boldsymbol{f}^{out}$ is ϵ_1-DP through a (non-trivial) application of the Output Perturbation algorithm by Chaudhuri et al. (2011). In the prediction phase, we must show that sufficient noise is added to ensure $\tilde{X}_{post}$ is an ϵ_2-DP version of X_{post}. Then privacy of $\boldsymbol{y}^{out}$ comes from the composition of these two private estimates.

We start by proving that $\boldsymbol{f}^{out}$ is ϵ_1-DP.

Theorem 3.6 *Step 1 of Algorithm 2 that computes* $\boldsymbol{f}^{out}$ *is* $(\epsilon_1, 0)$*-differentially private.*

Step 1 of Algorithm 2 instantiates the Laplace mechanism, and the crux of the proof lies in obtaining the sensitivity bound of the synthetic control query, which is fundamentally different from the setting of a traditional DP-ERM (Chaudhuri et al., 2011).

Lemma 3.7 *The* ℓ_2 *sensitivity of* $\boldsymbol{f}^{reg}$ *is* $\Delta\boldsymbol{f}^{reg} \leq \frac{4T_0\sqrt{8+n}}{\lambda}$.

The asymptotic dependence on n and T_0 may seem undesirable, but we show that it is unavoidable.

Lemma 3.8 *The* ℓ_2 *sensitivity of* $\boldsymbol{f}^{reg}$ *is* $\Delta\boldsymbol{f}^{reg} = \Omega(\sqrt{n})$.

Next we move to privacy of $\tilde{X}_{post}$ and its role in ensuring privacy of $\boldsymbol{y}^{out}$.

Lemma 3.9 *The computation of* $\tilde{X}_{post}$ *in Step 2 of Algorithm 2 is* $(\epsilon_2, 0)$*-differentially private.*

$\tilde{X}_{post}$ is privatized through another instantiation of the Laplace Mechanism by Dwork et al. (2006). Thus to prove Lemma 3.9, we only need to bound the sensitivity of X_{post}. We first note that the Frobenius norm of a matrix $X \in \mathbb{R}^{n\times(T-T_0)}$ is equal to the ℓ_2 norm of the equivalent flattened vector $X \in \mathbb{R}^{n(T-T_0)}$ (Horn and Johnson, 2012). Thus implementing the matrix-valued Laplace Mechanism with noise parameter calibrated to the ℓ_2 sensitivity of the flattened matrix-valued query over ϵ will ensure $(\epsilon, 0)$-differential privacy. Since all entries in X_{post} are bounded in $[-1, 1]$, each entry can change by at most 2 between two neighboring databases, which can differ in at most $T - T_0$) entries, hence the ℓ_2 sensitivity of flattened X_{post} is $2\sqrt{T-T_0}$.

Finally, we combine Theorem B.1 and Lemma 3.9 to complete the proof of Theorem 3.1. The estimates $\boldsymbol{f}^{out}$ and $\tilde{X}_{post}$ are together $(\epsilon_1+\epsilon_2, 0)$-differentially private by DP

composition, and then $\boldsymbol{y}^{out}$ is $(\epsilon_1+\epsilon_2, 0)$-differentially private by post-processing.

Since the prediction step of Algorithm 3 is identical to that of Algorithm 2, we only need to show that $\mathbf{f}^{obj}$ is computed in an (ϵ_1, δ)-DP manner (Theorem D.1) to complete the proof of Theorem 3.4.

Theorem 3.10 *Step 1 of Algorithm 3 that computes $\boldsymbol{f}^{obj}$ is (ϵ_1, δ)-differentially private.*

We show a proof sketch here and defer the full proof to Appendix D. At a high-level, the privacy of $\boldsymbol{f}^{obj}$ comes from a carefully modified instantiation of the Objective Perturbation algorithms by Chaudhuri et al. (2011) and Kifer et al. (2012). The Objective Perturbation method modifies the standard Ridge Regression objective function $J(\boldsymbol{f})$ by adding an additional regularization term and a noise term to ensure privacy:

$$\begin{aligned} J^{obj}(\boldsymbol{f}) &= J(\boldsymbol{f}) + \tfrac{\Delta}{2T_0}||\boldsymbol{f}||_2^2 + \tfrac{1}{T_0}\boldsymbol{b}^\top \boldsymbol{f} \\ &= \mathcal{L}(\boldsymbol{f}) + \tfrac{\lambda+\Delta}{2T_0}||\boldsymbol{f}||_2^2 + \tfrac{1}{T_0}\boldsymbol{b}^\top \boldsymbol{f}, \end{aligned}$$

where $\boldsymbol{b}$ is a random vector drawn from a high-dimensional Laplace distribution($\delta = 0$) or a multivariate Gaussian distribution($\delta > 0$).

Notice that $J^{obj}(\boldsymbol{f})$ is strongly convex (for any $\Delta \geq 0$) and differentiable. Hence, for any given input dataset $\mathcal{D} = \{X_{pre}, y_{pre}\}$ and any fixed parameters $(\lambda, \epsilon_1, \epsilon_2, \delta)$, there exists a bijection between a realized value of the noise term $\boldsymbol{b}$ and $\boldsymbol{f}^{obj} := \arg\min_{\boldsymbol{f}} J^{obj}(\boldsymbol{f})$ given that realized $\boldsymbol{b}$.[1] Let $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ be noise value that must have been realized when database $\mathcal{D}$ was input and $\boldsymbol{f}^{obj} = \boldsymbol{\alpha}$ was the output. We can then use this bijection to analyze the distribution over outputs on neighboring databases via the (explicitly given) noise distribution. Then, we can express the ratio between the two distributions over $\boldsymbol{f}^{obj}$ trained from neighboring databases, and tune the amount of noise depending on the privacy budget.

The ratio is a product of two terms, $\Gamma(\boldsymbol{\alpha})$ and $\Phi(\boldsymbol{\alpha}; \Delta)$. The parameter Δ serves a role to divide the ϵ_1 budget between these two terms, by distinguishing between two cases. In the first case, ϵ_1 is large enough that we can choose $\Delta = 0$ and still have some privacy budget (ϵ_0) remaining to bound $\Gamma(\boldsymbol{\alpha})$. In the other case, if ϵ_1 is too small to bound $\Phi(\boldsymbol{\alpha}; \Delta)$ with $\Delta = 0$, then we divide the privacy budget equally between bounding $\Gamma(\boldsymbol{\alpha})$ and $\Phi(\boldsymbol{\alpha}; \Delta)$, and find an appropriate value for $\Delta > 0$. For both cases, we prove that $\Phi(\boldsymbol{\alpha}; \Delta)$ is upper bounded by $e^{\epsilon_1 - \epsilon_0}$ (Lemma D.2) and $\Gamma(\boldsymbol{\alpha})$ is upper bounded by e^{ϵ_0} (Lemma D.5). Product of the two upper bounds provides the upper bound of the ratio, completing the proof of Theorem D.1.

[1] For a simple analogy, consider the one-dimensional Laplace Mechanism on query f and database x, which outputs $y = f(x) + Lap(\Delta f/\epsilon)$. Given f and x, there is a bijection between noise terms and outputs since the noise term must equal $y - f(x)$.

3.5 Proof Sketch for Accuracy Guarantees

Theorem 3.2 and 3.5 presents accuracy guarantee of the two algorithms in terms of root mean squared error (RMSE), defined as follows: $RMSE(\boldsymbol{y}^{out}) = \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||\boldsymbol{y}^{out} - \boldsymbol{m_{post}}||_2]$. We note that while it may seem most natural to bound the difference between $\boldsymbol{y}^{out}$ and $\boldsymbol{y}_{post}$, we instead use $\boldsymbol{m_{post}}$ for two reasons. Firstly, $\boldsymbol{y}_{post}$ may not match $\boldsymbol{y}_{pre}$ due to the intervention. Secondly, $\boldsymbol{m_{post}}$ captures the true signal that we are trying to estimate.

The proof aims to bound the expectation using the submultiplicative norm property (i.e., for any matrix A and vector $\boldsymbol{x}$, $||A\boldsymbol{x}||_2 \leq ||A||_2||\boldsymbol{x}||_2 \leq ||A||_F||\boldsymbol{x}||_2$) and the known distributions of noise terms $\boldsymbol{v}$ (Step 1 of Algorithm 2), $\boldsymbol{b}$ (Step 1 of Algorithm 3), and W (Step 2 of both algorithms). For example, we can decompose $\boldsymbol{y}^{out} = \tilde{X}_{post}^\top \boldsymbol{f}^{out} = (M_{post}^\top + Z^\top + W^\top)(\boldsymbol{f}^{reg} + \boldsymbol{v})$, and compare against $\boldsymbol{m_{post}} = M_{post}^\top \boldsymbol{f}$.

The full proof of Theorem 3.2 and 3.5 are presented in Appendices C and E.

4 EMPIRICAL PERFORMANCE

This section presents the empirical performance of both $DPSC_{out}$ and $DPSC_{obj}$ on synthetic datasets. We create four synthetic datasets following the modeling assumptions of Section 2.1, with the number of pre-intervention observations $T_0 \in \{10, 100\}$ and the number of donors $n \in \{10, 100\}$. We set $T = T_0 + 3$, meaning that the performance will be measured by RMSE of the next three data points. More details about the data generation process can be found in Appendix A.1

For a fair comparison, we use $\delta = 0$ for the objective perturbation, so the only privacy parameter we consider is $\epsilon = \epsilon_1 + \epsilon_2$. We show two experiments, one with fixed ϵ and varying λ, the other with fixed and varying ϵ. Each experiment was repeated 500 times on each dataset, and the error bands in all figures show 95% confidence intervals taken over the randomness in the algorithms. We present four graphs highlighting the main findings and defer all detailed plots and additional analyses to Appendix A.

4.1 Results

For the first set of experiments, we fix $\epsilon_1 = \epsilon_2 = 50$ and vary the regularization parameter $\lambda \in \{5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$. The top two graphs in Figure 3 show that the optimal choice of λ roughly remains $\lambda = T_0$ for all sizes of database considered. The orange and blue curves in the figure corresponding to the two databases with $T_0 = 10$ are approximately minimized at $\lambda = 10$, while the green and red curves with

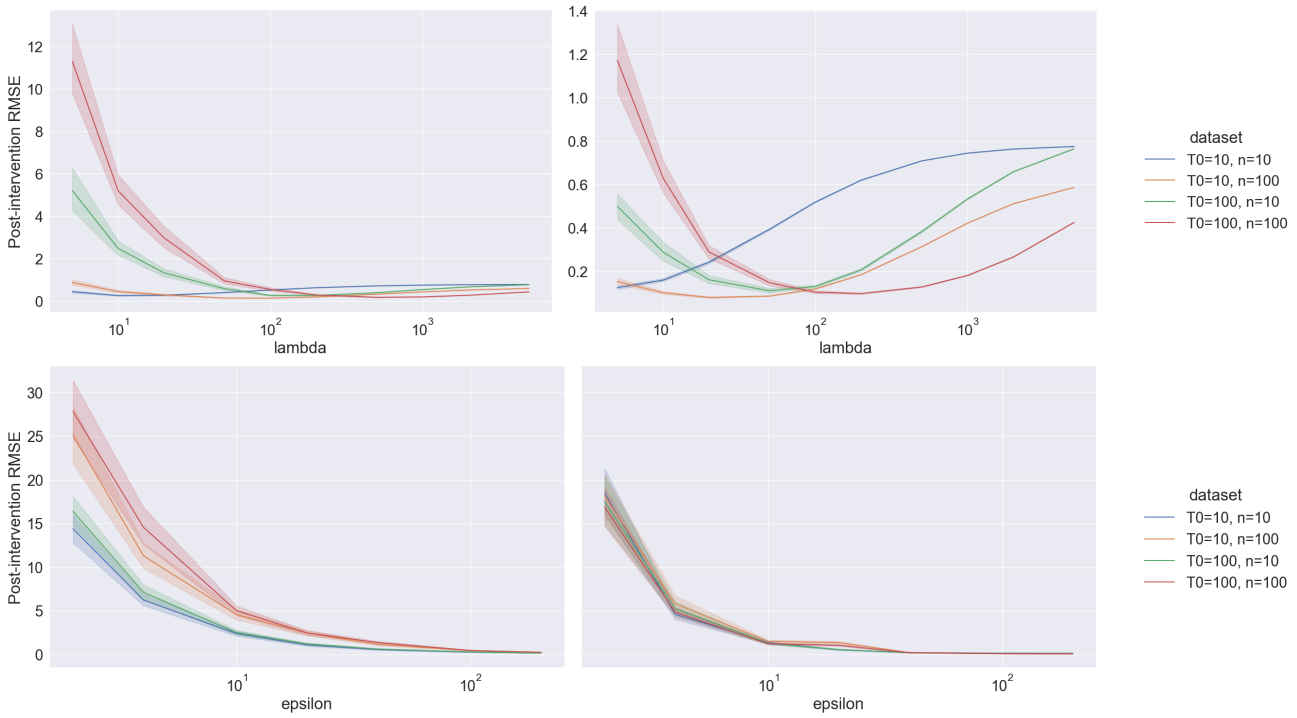

Figure 2: Post-intervention RMSE of $DPSC_{out}$ (left) and $DPSC_{obj}$ (right) as a function of λ (top) and as a function of ϵ (bottom) on four synthetic datasets of varying size.

$T_0 = 100$ are approximately minimized at $\lambda = 100$. While this is clearer in the figure for Objective Perturbation, it is also true for Output Perturbation, although the U-shape is less visible due to the larger scale of the y-axis.

Additionally, we observe that the RMSE of the private methods is larger than that of the non-private method for smaller $\lambda \leq 20$, with objective perturbation substantially outperforming output perturbation. Also, the empirical performance of both algorithms is dramatically better than the theoretical bounds may suggest, which implies potential room for improvements in the bound of Theorem 3.5. A more detailed discussion can be found in Appendix A.2.

For the next set of experiments, we fix $\lambda = T_0$ and run more experiments with varying $\epsilon \in \{2, 4, 10, 20, 40, 100, 200\}$, where the privacy budget is split equally between Step 1 and Step 2 (i.e., $\epsilon_1 = \epsilon_2 = \epsilon/2$.) The bottom two plots in Figure 2 show that the RMSE diminishes as ϵ grows, as expected. We continue to observe $DPSC_{obj}$ outperforming $DPSC_{out}$ for most ϵ values. At $\epsilon = 10$ for $DPSC_{out}$, we see that RMSE on datasets with $n = 100$ (red and orange lines) is higher than that of databases with $n = 10$ (blue and green lines). This is consistent with our theoretical analysis in Section C.1.1 that the RMSE of $DPSC_{out}$ is $O(\frac{n}{\epsilon})$. For $DPSC_{obj}$, the accuracy bound has an additional dependency on T_0 (Corollary E.2), so the ordering of performance by database size is less clear.

4.2 Guidance for Hyperparameter Tuning

Both the $DPSC_{out}$ and $DPSC_{obj}$ algorithms require tuning the hyperparameter λ. This parameter plays an important role in determining the amount of noise, since it appears in the sensitivity of $\boldsymbol{f}$, i.e., $\Delta \boldsymbol{f} = \frac{4T_0\sqrt{8+n}}{\lambda}$. In theory, the optimal choice of this parameter is recommended to be $\lambda = O(T_0)$ because the regression coefficient in the objective function is $\frac{\lambda}{2T_0}$, and the importance of the regularizer should not diminish as T_0 increases (i.e., as we have more training data points). We confirm this empirically by plotting the post-intervention RMSE as a function of λ (Figure 2, top), which is near-optimal around $\lambda = T_0$.

5 Conclusion

This paper is the first to propose differentially private versions of the synthetic control algorithm. We provide algorithms based on output perturbation and objective perturbation, and provide formal privacy and accuracy guarantees for each. Our main technical contribution is a novel analysis of the sensitivity of regression in the *transposed* setting, which also impacted our accuracy analysis. To enable practical use of the new tools, we provide a closed-form accuracy bound for both algorithms under distributional assumptions and guidance to practitioners for tuning the parameters of each algorithm.

Acknowledgements

The first and third authors were supported in part by Novartis AG. The first and second author was supported in part by NSF grant CNS-1942772 (CAREER), a Mozilla Research Grant, a JPMorgan Chase Faculty Research Award, and an Apple Privacy-Preserving Machine Learning Award.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132.

Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, pages 1–18.

Agarwal, A., Misra, V., Shah, D., Shen, D., 1, A., and 2, A. (2021a). Identifying personalized treatment effect from population-level clinical trials data for alzheimer's patient. *Working Paper*.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021b). On robustness of principal component regression. *Journal of the American Statistical Association*, (just-accepted):1–34.

Amjad, M., Misra, V., Shah, D., and Shen, D. (2019). mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–27.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, pages 1–15.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

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

Ben-Michael, E., Feller, A., and Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, (just-accepted):1–34.

Billingsley, P. (1995). Measure and probability.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

Cummings, R., Ioannidis, S., and Ligett, K. (2015). Truthful linear regression. In *Conference on Learning Theory*, pages 448–483. PMLR.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Dube, A. and Zipperer, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284.

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., Doubek, M., Fielding, A. K., Giebel, S., Haddad, V., et al. (2016). Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood cancer journal*, 6(9):e473–e473.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, USA, 2nd edition.

Kearns, M., Pai, M., Roth, A., and Ullman, J. (2014). Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 403–410. ACM.

Kifer, D., Smith, A., and Thakurta, A. (2012). Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings.

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., and Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics*, 25(12):1514–1528.

Petrone, J. (2018). Roche pays $1.9 billion for flatiron's army of electronic health record curators. *Nature Biotechnology*, 36(4):289–291.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Thorlund, K., Dron, L., Park, J., and Mills, E. (2020). Synthetic and external controls in clinical trials – a primer for researchers. *Clinical Epidemiology*, Volume 12:457–467.

Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

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

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, D., Ye, M., and Xu, J. (2017). Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30.

A Experimental Settings and Results

In this section, we provide more details about the experiments that were omitted from the main text due to the page limit.

A.1 Synthetic Data Generation

We use synthetic datasets in our experiments, which enables us to observe the impact of varying the relevant parameters in the data and to match the modeling assumptions of Section 2.1. We create use $T_0 \in \{10, 100\}$ and $n \in \{10, 100\}$, corresponding to both smaller and larger number of donors and observations, and we always use $T = T_0 + 3$, meaning that the synthetic control algorithm must predict the next three data points.

The true signals M and $\mathbf{m}$ are generated according to a linear model with random slope, formalized as:

$$M_{i,t} = \theta_i t \quad \text{and} \quad m_t = \theta_0 t, \ \forall i \in [n], t \in [T],$$

where the θ_i are sampled i.i.d. from a truncated Gaussian with mean 4, variance 1, and support $[3, 5]$. Elements of the noise terms Z and $\mathbf{z}$ are sampled i.i.d. from a truncated Gaussian with mean zero, variance 0.1 and support $[-1, 1]$. Following Equation 1, the donor and target data were respectively $X = M + Z$ and $\mathbf{y} = \mathbf{m} + \mathbf{z}$. Figure 3 shows an example synthetic dataset generated in this way, with the donor data in grey and the target in red.

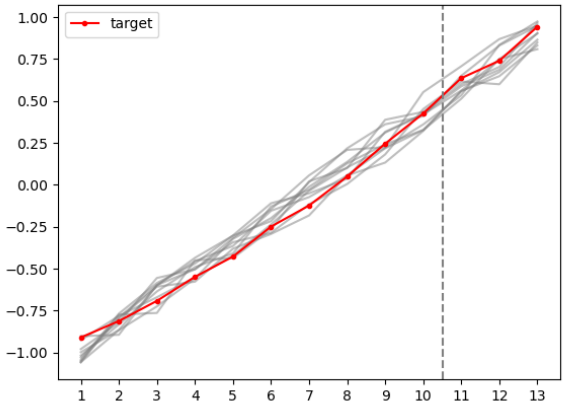

Figure 3: Illustration of example synthetic dataset generated with $T_0 = 10$ and $n = 10$. The target time series is in red, and the donor time series are all in grey.

In each experiment with a fixed T_0 and n, a single database was generated, and then algorithms were run 500 times on each dataset. We evaluate post-intervention RMSE as the accuracy measure of interest, as in our theoretical results. Error bands in all figures show 95% confidence intervals, taken over the randomness in the algorithms.

A.2 Optimizing regularization parameter λ

The first question we aim to address in our experiments is the impact of the parameter λ on performance, and guidance for analysts in their choice of optimal λ. In our first set of experiments, we fixed $\epsilon_1 = \epsilon_2 = 50$, $T_0 = 10$, and $n = 10$—other values of ϵ and (T_0, n) are considered respectively in Sections A.3 and **??**—and empirically measured pre- and post-intervention RMSE as a function of λ.

Figure 4 shows the post-intervention RMSE of $DPSC_{out}$, $DPSC_{obj}$, and non-private synthetic control as a function of λ, for values $\lambda \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$.

We observe that the performance of the three methods converges as λ grows large, but that the RMSE of the private methods is larger than that of the non-private method for smaller $\lambda \leq 20$, with Objective Perturbation substantially outperforming Output Perturbation.

To aid the analyst in choosing an optimal λ, we observe that the RMSE of DPSC is minimized around $\lambda = T_0$ for all four datasets (see Figure 2 in Section 4.1). This is consistent with our theoretical recommendations that λ should be $O(T_0)$.

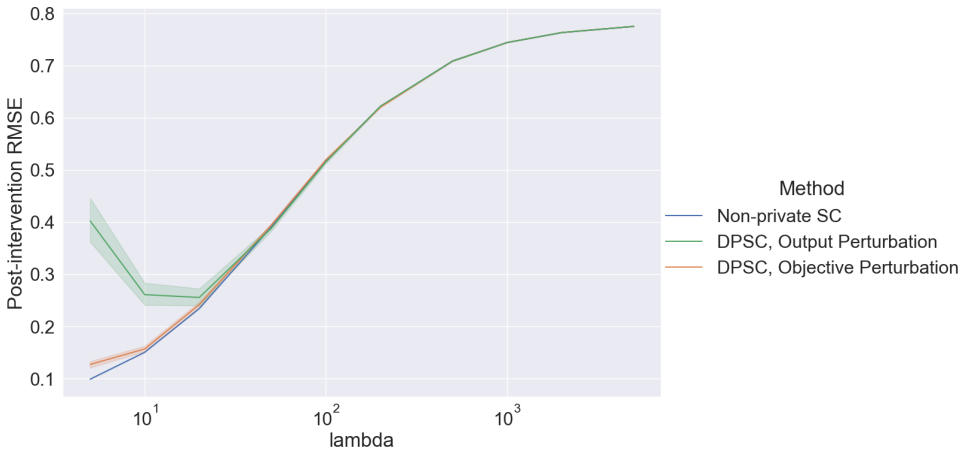

Figure 4: Behavior of post-intervention RMSE over λ, tested on a synthetic datast with $T_0 = 10, n = 10$ for the synthetic control methods of non-private SC (blue), $DPSC_{out}$ (green), and $DPSC_{obj}$ (orange).

The U-shape has a natural theoretical explanation: smaller λ increases sensitivity and thus privacy noise and RMSE, while larger λ increases the weight of the regularization term in the loss function, which will cause all three regularized methods to converge to each other.

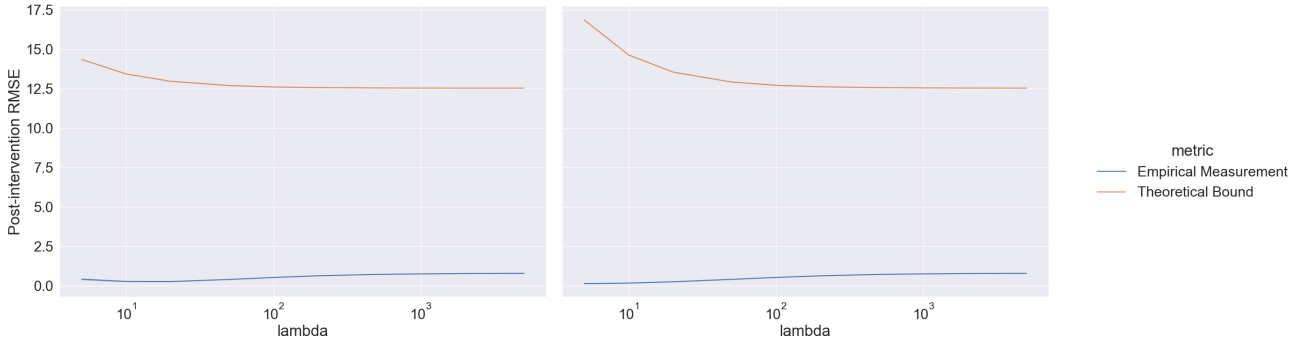

Figure 5: Comparison of post-intervention RMSE in theory versus in practice, using $DPSC_{out}$ (left) and $DPSC_{obj}$ (right) on a dataset of size $n = 10, T_0 = 10$.

Figure 5 compares the empirical post-intervention RMSE of $DPSC_{out}$ and $DPSC_{obj}$ with the theoretical guarantees of Theorem 3.2 and 3.5 instantiated with parameters of our experiments. We observe that the empirical performance of both algorithms is dramatically better than the theoretical bounds may suggest. We also observe that $DPSC_{obj}$ (right) has lower empirical error that $DPSC_{out}$ (left), which diverges from our theoretical predictions. This suggests potential room for theoretical improvements in the bound of Theorem 3.5.

A.3 Effect of privacy paramter ϵ

Next, we address the effect of ϵ in the performance of both $DPSC_{out}$ and $DPSC_{obj}$. In these experiments, we use $\lambda = T_0$ based on the findings in Section A.2 and consider overall privacy budget $\epsilon = \epsilon_1 + \epsilon_2$ with $\epsilon_1 = \epsilon_2 = \epsilon/2$. That is, the privacy budget is split evenly between the regression and projection steps in both algorithms. Results are presented for $\epsilon \in \{2, 4, 10, 20, 40, 100, 200\}$; stronger privacy guarantees (i.e., $\epsilon \leq 2$) were tested but excluded from the plots due to substantially higher RMSE values.

Figure 6 shows the post-intervention RMSE of $DPSC_{out}$ and $DPSC_{obj}$. As is to be expected, error diminishes with larger ϵ. We also continue to observe $DPSC_{obj}$ outperforming $DPSC_{out}$ for most ϵ values, as in Section A.2. $DPSC_{out}$ performs slightly better thatn $DPSC_{obj}$ at $\epsilon = 2$ in this dataset ($T_0 = 10$ and $n = 10$); however, it is not the case for all datasets (See Figure 2 in Section 4.1).

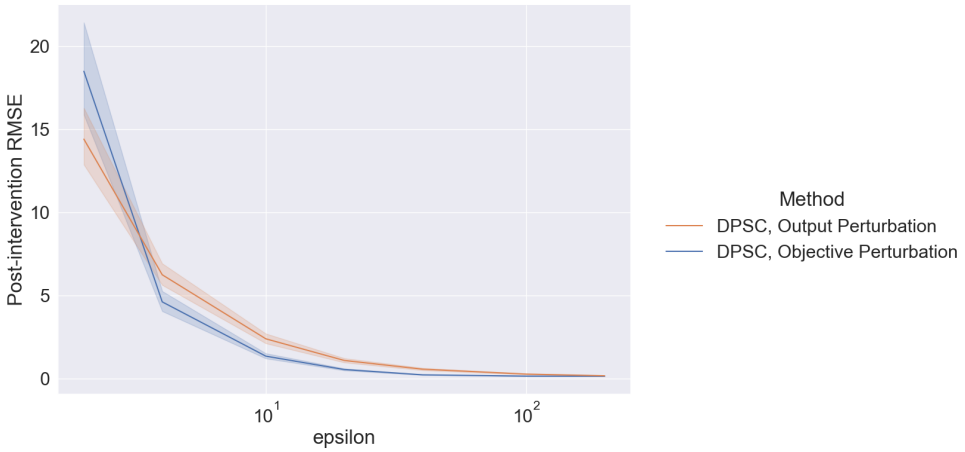

Figure 6: Post-intervention RMSE of $DPSC_{out}$ (blue), and $DPSC_{obj}(red)$ for varying ϵ, tested on a dataset with $T_0 = 10$ and $n = 10$.

For epsilon-regimes that are closer to the values chosen in practice (i.e., $\epsilon \leq 4$), the empirical RMSE was too high for practical use. We suggest a few methods to remedy this in future work. First, a rejection sampling step can be introduced between the learning and projection steps of each algorithm that compares the noisy $\boldsymbol{f}^{out}$ and the original $\boldsymbol{f}^{reg}$. This step must also be done differentially privately to maintain the overall privacy guarantee. Additionally, our experiments only considered pure-DP with $\delta = 0$; relaxing to approximate-DP with $\delta > 0$ would likely yield lower RMSE.

B Privacy Guarantees of $DPSC_{out}$

In this section, we will prove Theorem 3.1, that DPSC is $(\epsilon_1 + \epsilon_2, 0)$-differentially private. The proof relies on the privacy of $\boldsymbol{f}^{out}$ in the learning phase, and then $\tilde{X}_{post}$ in the prediction phase. At a high level, $\boldsymbol{f}^{out}$ is ϵ_1-DP through a (non-trivial) application of the Output Perturbation algorithm of Chaudhuri et al. (2011). The non-triviality comes from the vertical regression used in synthetic control, rather than the horizontal regression classically used in empirical risk minimization (as illustrated in Figure 1), which requires novel sensitivity analysis of the function $\boldsymbol{f}^{reg}$. In the prediction phase, we must show that sufficient noise is added to ensure $\tilde{X}_{post}$ is an ϵ_2-DP version of X_{post}. Then privacy of $\boldsymbol{y}^{out}$ comes from post-processing and composition of these two private estimates.

B.1 Privacy of $\boldsymbol{f}^{out}$

Let us begin by proving that $\boldsymbol{f}^{out}$ is ϵ_1-DP.

Theorem B.1 *Step 1 of Algorithm 2 that computes $\boldsymbol{f}^{out}$ is $(\epsilon_1, 0)$-differentially private.*

It might seem that Theorem 3.1 should follow immediately from the privacy guarantees of Output Perturbation in Chaudhuri et al. (2011). Indeed, Theorem 6 of Chaudhuri et al. (2011) states that a similar algorithm is $(\epsilon, 0)$-DP under certain technical conditions. However, the proof of this result relies on sensitivity analysis of classical empirical risk minimization (see Corollary 8 of Chaudhuri et al. (2011)) which does not hold in the synthetic control setting. The crux of the difference comes from the vertical regression (i.e., along the columns) of synthetic control as illustrated in Figure 1, while privacy must still be maintained along the rows. Thus the sensitivity of $\boldsymbol{f}^{reg}$ to a change in a single *donor row* is fundamentally different from the sensitivity in a standard empirical risk minimization setting. See Remark B.5 for a more technical exploration of this difference. Additionally, while the ERM framework of Chaudhuri et al. (2011) is fully general, their results (including Theorem 6 and Corollary 8) apply only to the problem setting of binary classification via logistic regression, by assuming a specific loss function $\mathcal{L}$ in the analysis.

Instead, we prove Theorem B.1 primarily using first-principles (i.e., direct sensitivity analysis and the Laplace Mechanism of Dwork et al. (2006), which also underpins the results of Chaudhuri et al. (2011)) starting with Lemma 3.7. The proof of Lemma 3.7 and Theorem B.1 will be augmented with one intermediate result for output perturbation from Chaudhuri et al. (2011) that does apply to our setting, and one fact from Cummings et al. (2015), which extended the binary classification result of Chaudhuri et al. (2011) to the Ridge regression loss function that we use.

Lemma 3.7 *The ℓ_2 sensitivity of $\boldsymbol{f}^{reg}$ is $\Delta \boldsymbol{f}^{reg} \leq \frac{4T_0\sqrt{8+n}}{\lambda}$.*

To prove Lemma 3.7, we will first use the following lemma from Chaudhuri et al. (2011), which bounds the sensitivity of $\boldsymbol{f}^{reg}$ as a function of the strong convexity parameter of the loss function $\mathcal{L}$.

Lemma B.2 (Chaudhuri et al. (2011), Lemma 7) *Let $G(\boldsymbol{f})$ and $g(\boldsymbol{f})$ be two vector-valued functions, which are continuous, and differentiable at all points. Moreover, let $G(\boldsymbol{f})$ and $G(\boldsymbol{f}) + g(\boldsymbol{f})$ be λ-strongly convex. If $\boldsymbol{f}_1 = \arg\min_{\boldsymbol{f}} G(\boldsymbol{f})$ and $\boldsymbol{f}_2 = \arg\min_{\boldsymbol{f}} G(\boldsymbol{f}) + g(\boldsymbol{f})$, then*

$$\|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_2 \leq \frac{1}{\lambda} \max_{\boldsymbol{f}} \|\nabla g(\boldsymbol{f})\|_2.$$

We instantiate this lemma by defining

$$G(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}, \mathcal{D}) \quad \text{and} \quad g(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}, \mathcal{D}') - \mathcal{L}(\boldsymbol{f}, \mathcal{D}), \tag{5}$$

for two arbitrarily neighboring databases $\mathcal{D}, \mathcal{D}'$ and defining the following two maximizers:

$$\boldsymbol{f_1} = \arg\min \mathcal{L}(\boldsymbol{f}, \mathcal{D}) = \arg\min G(\boldsymbol{f}) \quad \text{and} \quad \boldsymbol{f_2} = \arg\min \mathcal{L}(\boldsymbol{f}, \mathcal{D}') = \arg\min G(\boldsymbol{f}) + g(\boldsymbol{f}).$$

Then,

$$\Delta \boldsymbol{f}^{reg} = \max_{D, D' \text{ neighbors}} \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_2.$$

To apply Lemma B.2, we must show that $G(\boldsymbol{f})$ and $g(\boldsymbol{f})$ are continuous and differentiable. $G(\boldsymbol{f})$ is simply the Ridge regression loss function, which is known to be continuous and differentiable Hastie et al. (2009). Since $g(\boldsymbol{f})$ is the difference between two continuous and differentiable functions, then it is also continuous and differentiable Boyd and Vandenberghe (2004). We must also show strong convexity of $G(\boldsymbol{f})$ and $G(\boldsymbol{f}) + g(\boldsymbol{f})$. The following lemma from Cummings et al. (2015) immediately gives that these two functions are both λ-strongly convex.

Lemma B.3 (Cummings et al. (2015), Lemma 32) *The Ridge regression loss function with regularizer $\frac{\lambda}{2T_0}$ is $\frac{\lambda}{T_0}$-strongly convex.*

Thus by Lemma B.2, the sensitivity $\Delta \boldsymbol{f}^{reg} = \max_{D, D' \text{ neighbors}} \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_2 \leq \frac{T_0}{\lambda} \max_{\boldsymbol{f}} \|\nabla g(\boldsymbol{f})\|_2$. All that remains is to bound $\|\nabla g(\boldsymbol{f})\|_2$. A proof of the following lemma is deferred to Appendix F.

Lemma B.4 *Let $g(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}, \mathcal{D}') - \mathcal{L}(\boldsymbol{f}, \mathcal{D})$ for two arbitrarily neighboring databases $\mathcal{D}, \mathcal{D}'$. Then,*

$$\max_{\boldsymbol{f}} \|\nabla g(\boldsymbol{f})\| \leq 4\sqrt{8+n}.$$

Remark B.5 *If we were instead considering simple linear regression in the classical setting (i.e., as in Chaudhuri et al. (2011)) using T_0 data points with n dimensional features, $g(\boldsymbol{f})$ would only contain one term in the error, namely, the one data point $(\boldsymbol{x_i}, y_i)$ that differed across two neighboring databases. This yields*

$$g(\boldsymbol{f}) = \frac{1}{T_0}((\boldsymbol{x_i'} - \boldsymbol{x_i})^\top \boldsymbol{f} - (y_i' - y_i))^2$$

with gradient

$$\nabla g(\boldsymbol{f}) = \frac{2}{T_0}((\boldsymbol{x_i'} - \boldsymbol{x_i})(\boldsymbol{x_i'} - \boldsymbol{x_i})^\top \boldsymbol{f} - (y_i' - y_i)(\boldsymbol{x_i'} - \boldsymbol{x_i})),$$

which can be bounded by $O(\frac{1}{T_0})$. This result does not depend on the dimension of the features (n) and only depends on the number of data points (T_0).

However, in synthetic control, terms do not cancel as neatly across neighboring databases, and instead,

$$g(\boldsymbol{f}) = \frac{1}{T_0} \sum_{t=1}^{T_0} \left[\left(\boldsymbol{x_t}^\top \boldsymbol{f} - y_t \right) + \left(\boldsymbol{x_t}^\top \boldsymbol{f} - x_{i,t} f_i + x'_{i,t} f_i - y_t \right) \right] (x'_{i,t} - x_{i,t}) f_i.$$

Through a more involved analysis of this expression, we get the bound of Lemma B.4, which depends on n, rather than T_0.

Using these lemmas, we can now bound the sensitivity of our query, to complete the proof of Lemma 3.7.

$$\Delta \boldsymbol{f}^{reg} = \max_{\mathcal{D},\mathcal{D}' \text{ neighbors}} ||\boldsymbol{f}(\mathcal{D}) - \boldsymbol{f}(\mathcal{D}')||_2 \leq \max ||\boldsymbol{f_1} - \boldsymbol{f_2}|| \leq \frac{4T_0\sqrt{8+n}}{\lambda}. \tag{6}$$

Theorem B.1 then follows from the privacy guarantee of the high-dimensional Laplace Mechanism instantiated with the appropriate sensitivity value.

B.1.1 Dependence on n

One might wonder whether the asymptotic dependence on n and T_0 in the sensitivity is necessary. In practice, one should set $\lambda = O(T_0)$ (as discussed in greater detail in Section C.1.1), so the dependence on T_0 will not affect the accuracy of the algorithm. However, as we show next in Lemma 3.8, the dependence on n is asymptotically tight.

Lemma 3.8 *The ℓ_2 sensitivity of $\boldsymbol{f}^{reg}$ is $\Delta \boldsymbol{f}^{reg} = \Omega(\sqrt{n})$.*

Consider two neighboring databases (X, y) and (X', y), where $\boldsymbol{y} = \mathbf{1} \in \mathbb{R}^{T_0}$, $X \in \mathbb{R}^{n \times T_0}$ has all entries $1/n$, except the first row, which is all 1s. Neighboring database X' differs from X only in the first row, which is instead all 0s, and all other entries and $1/n$. The dimensions in this example are chosen to be $T_0 = n$, and we choose $\lambda = 2T_0$, so that the regularization coefficient is 1.

Computing the minimizers of the loss functions under each neighboring database using the closed-form expression yields $\boldsymbol{f}^{reg} = (XX^\top + I)^{-1}X\boldsymbol{y}$ with the first coordinate equal to $\frac{n^2}{n^2+2n-1}$, and all other coordinates are $\frac{n}{n^2+2n-1}$, and $\boldsymbol{f}^{reg\prime} = (X'X'^\top + I)^{-1}X'\boldsymbol{y}$ with first coordinate 0 and all other coordinates $\frac{-n}{1-2n}$. This yields ℓ_2 difference of,

$$||\boldsymbol{f}^{reg} - \boldsymbol{f}^{reg\prime}||_2 = \sqrt{\left(\frac{n^2}{n^2+2n-1}\right)^2 + (n-1)\left(\frac{n^3}{(n^2+2n-1)(1-2n)}\right)^2} = \Theta(\sqrt{n}).$$

Since we have a pair of neighboring databases with ℓ_2 distance in their output of $\Theta(\sqrt{n})$, then the sensitivity of $\boldsymbol{f}^{reg}$ cannot be $o(\sqrt{n})$.

Remark B.6 *We note that while the example in Lemma 3.8 is mathematically valid, such a degenerate case where all the donors are identical except for one person and the (exact) rank of the donor matrix is 1 is unlikely to happen in practical settings. Thus suggests that with additional domain knowledge on the selection criteria for donors, practitioners may be able to reduce the sensitivity and thus add less noise for privacy in special restricted cases of interest.*

B.2 Privacy of $\tilde{X}_{post}$ and $\boldsymbol{y}^{out}$

Next we move to privacy of $\tilde{X}_{post}$ and its role in ensuring privacy of $\boldsymbol{y}^{out}$.

Lemma 3.9 *The computation of $\tilde{X}_{post}$ in Step 2 of Algorithm 2 is $(\epsilon_2, 0)$-differentially private.*

$\tilde{X}_{post}$ is privatized through a simple application of the Laplace Mechanism of Dwork et al. (2006). Thus to prove Lemma 3.9, we need only to bound the sensitivity of X_{post} to show that the algorithm adds sufficient noise. We first note that the Frobenius norm of a matrix $X \in \mathbb{R}^{n \times (T-T_0)}$ is equal to the ℓ_2 norm of the equivalent flattened vector $X \in \mathbb{R}^{n(T-T_0)}$ Horn and Johnson (2012). Thus implementing the matrix-valued Laplace Mechanism with noise parameter calibrated to the ℓ_2 sensitivity of the flattened matrix-valued query over ϵ will ensure $(\epsilon, 0)$-differential privacy.

Lemma B.7 *The ℓ_2 sensitivity of flattened X_{post} is $2\sqrt{(T-T_0)}$.*

Changing one donor unit in X_{post} can change at most $T - T_0$ entries in the matrix. Since all entries in X_{post} are bounded in $[-1, 1]$, each data point can change by at most 2 between two neighboring databases. Thus viewing X_{post} as a flattened matrix, this will change the ℓ_2-norm of X_{post} by at most $2\sqrt{(T-T_0)}$.

Finally, we can combine Theorem B.1 and Lemma 3.9 to complete the proof of Theorem 3.1. The estimates $\boldsymbol{f}^{out}$ and $\tilde{X}_{post}$ are together $(\epsilon_1 + \epsilon_2, 0)$-differentially private by DP composition, and then $\boldsymbol{y}^{out}$ is $(\epsilon_1 + \epsilon_2, 0)$-differentially private by post-processing. We note that if one wanted to publish $\boldsymbol{f}^{out}$, this would not incur any additional privacy loss.

C Accuracy Guarantees of $DPSC_{out}$

In this section we will analyze the accuracy of $DPSC_{out}$. We first prove Theorem 3.2, restated below for convenience.

Theorem 3.2 *The estimator $\boldsymbol{y}^{out}$ output by Algorithm 2 satisfies:*

$$RMSE(\boldsymbol{y}^{out}) \leq \frac{||M_{post}||_2}{\sqrt{T-T_0}}\left(\mathbb{E}[||\boldsymbol{f}^{reg}-\boldsymbol{f}||_2]+\frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right)+\left(\sqrt{n\sigma^2}+\frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi+\frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right),$$

where $||\boldsymbol{f}^{reg}||_\infty \leq \psi$ for some $\psi > 0$, and RMSE is the root mean squared error of the estimator, defined as $RMSE(\boldsymbol{y}^{out}) = \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||\boldsymbol{y}^{out}-\boldsymbol{m}_{post}||_2]$.

This theorem gives bounds on the predicted post-intervention target vector $\boldsymbol{y^{out}}$, as measured by RMSE. This result is stated in full generality with respect to the distribution of data and the latent variables, and thus the bound depends on terms such as $||M_{post}||_{2,2}$ and $\mathbb{E}[||\boldsymbol{f}^{reg}-\boldsymbol{f}||_2]$. This is consistent with comparable bounds on the RMSE of robust synthetic control Amjad et al. (2018) which also depended on these terms (although the stated bounds of Amjad et al. (2018) suppress dependence on n). Section C.1 provides a proof of this main result.

Analysts may still wonder about the full asymptotic performance of $DPSC_{out}$ algorithm. To this end, in Section C.2, we additionally derive closed-form bounds for these distribution-dependent terms (under some mild assumptions). We present Corollary C.5, which gives a bound on RMSE of $\boldsymbol{y}^{out}$ that depends only on input parameters of the algorithm and the model.

C.1 Accuracy of post-intervention prediction $\boldsymbol{y}^{out}$

We will prove Theorem 3.2 by showing that the prediction vector $\boldsymbol{y}^{out}$ output by $DPSC_{out}$ in Algorithm 2 is close to the true values, as measured by Root Mean Squared Error (RMSE), defined as follows:

$$RMSE(\boldsymbol{y}^{out}) = \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||\boldsymbol{y}^{out}-\boldsymbol{m}_{post}||_2]. \tag{7}$$

We note that while it may seem most natural to bound the difference between $\boldsymbol{y}^{out}$ and $\boldsymbol{y}_{post}$, we instead use $\boldsymbol{m}_{post}$ for two reasons. Firstly, $\boldsymbol{y}_{post}$ may not even match $\boldsymbol{y}_{pre}$ due to the intervention. Secondly, $\boldsymbol{m}_{post}$ captures the true signal that we are trying to estimate, which is the counterfactual outcome without the intervention.

We begin by bounding the expected ℓ_2 difference between $\boldsymbol{y}^{out}$ and $\boldsymbol{m}_{post}$. Using the fact that

$$\boldsymbol{y}^{out} = \tilde{X}_{post}^\top \boldsymbol{f}^{out} = (X_{post}^\top + W^\top)(\boldsymbol{f}^{reg}+\boldsymbol{v}),$$

and that $X_{post} = M_{post} + Z$ and $\boldsymbol{m} = M_{post}^\top \boldsymbol{f}$ (by Equation (2)), we can expand the expectation as follows:

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{y}^{out}-\boldsymbol{m}_{post}||_2] &= \mathbb{E}[||(X_{post}^\top+W^\top)(\boldsymbol{f}^{reg}+\boldsymbol{v})-M_{post}^\top\boldsymbol{f}||_2] \\
&= \mathbb{E}[||(M_{post}^\top+Z^\top+W^\top)(\boldsymbol{f}^{reg}+\boldsymbol{v})-M_{post}^\top\boldsymbol{f}||_2] \\
&\leq \mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg}-\boldsymbol{f})||_2+||(Z^\top+W^\top)\boldsymbol{f}^{reg}||_2+||(M_{post}^\top+Z^\top+W^\top)\boldsymbol{v}||_2] \\
&= \mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg}-\boldsymbol{f})||_2]+\mathbb{E}[||(Z^\top+W^\top)\boldsymbol{f}^{reg}||_2]+\mathbb{E}[||(M_{post}^\top+Z^\top+W^\top)\boldsymbol{v}||_2]
\end{aligned} \tag{8}$$

We next proceed to bound each of the terms in Equation (8) separately, making use of the following submultiplicative norm property, which holds for any matrix A and vector $\boldsymbol{x}$:

$$||A\boldsymbol{x}||_2 \leq ||A||_2||\boldsymbol{x}||_2 \leq ||A||_F||\boldsymbol{x}||_2, \tag{9}$$

where $||A||_2 = ||A||_{2,2}$ is the spectral norm of A, $||A||_F$ is the Frobenius norm of A, and $||\boldsymbol{x}||_2$ is the ℓ_2 norm of $\boldsymbol{x}$.

We also know the distribution of the norms of noise terms $\boldsymbol{v}$ and W that were added to preserve privacy, because they were constructed explicitly within Algorithm 2:

$$\mathbb{E}[||\boldsymbol{v}||_2] = \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1} \quad \text{and} \quad \mathbb{E}[||W||_F] = b = \frac{2\sqrt{T-T_0}}{\epsilon_2}. \tag{10}$$

Using these facts, we can obtain bounds for the three terms in (8). A complete proof of Lemma C.1 can be found in Appendix F.2.

Lemma C.1 *The three terms in Equation* (8) *can be bounded as follows:*

$$\mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg} - \boldsymbol{f})||_2] \leq ||M_{post}||_{2,2} \cdot \mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2],$$

$$\mathbb{E}[||(Z^\top + W^\top)\boldsymbol{f}^{reg}||_2] \leq \sqrt{n}\psi\left(\sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2}\right), \text{ and}$$

$$\mathbb{E}[||(M_{post}^\top + Z^\top + W^\top)\boldsymbol{v}||_2] \leq \left(||M_{post}||_{2,2} + \sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2}\right)\frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}.$$

Applying the bounds of Lemma C.1 to Equation (8) yields,

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{y}^{out} - \boldsymbol{m}||_2] &\leq ||M_{post}||_{2,2} \cdot \mathbb{E}[\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] + \sqrt{n}\psi\left(\sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2}\right) \\
&\quad + \left(||M_{post}||_{2,2} + \sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2}\right)\frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1} \\
&\leq ||M_{post}||_{2,2}\left(\mathbb{E}[\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] + \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right) \\
&\quad + \left(\sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right)
\end{aligned}$$

Combining this with Equation (7) gives the desired bound for Theorem 3.2:

$$RMSE(\boldsymbol{y}^{out}) \leq \frac{||M_{post}||_{2,2}}{\sqrt{T-T_0}}\left(\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] + \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right) + \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right).$$

C.1.1 Cost of privacy in synthetic control

To understand the additional error incurred due to privacy, compare the bound of Theorem 3.1 to the RMSE of the equivalent non-private prediction, $\boldsymbol{y}^{reg} = X_{post}^\top \boldsymbol{f}^{reg}$.

$$\begin{aligned}
RMSE(\boldsymbol{y}^{reg}) &= \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||X_{post}^\top \boldsymbol{f}^{reg} - \boldsymbol{m}||_2] \\
&= \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||(M_{post}^\top + Z_{post}^\top)\boldsymbol{f}^{reg} - M_{post}^\top \boldsymbol{f}||_2] \\
&\leq \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg} - \boldsymbol{f})||_2 + ||Z_{post}^\top \boldsymbol{f}^{reg}||_2] \\
&\leq \frac{||M_{post}||_{2,2}}{\sqrt{T-T_0}}\left(\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2]\right) + \sqrt{n}\psi \cdot \sqrt{n\sigma^2} \qquad (11)
\end{aligned}$$

Lemma C.6 in the next section shows that $\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] = O(\sqrt{n})$. Then the first term of Equation (11) can be easily bounded using the following fact,

$$||M_{post}||_{2,2} \leq ||M_{post}||_F \leq \sqrt{n(T-T_0)},$$

so $\frac{||M_{post}||_{2,2}}{\sqrt{T-T_0}} \leq \sqrt{n}$. Thus we see that $RMSE(\boldsymbol{y}^{reg}) = O(n)$.

Comparing Equation (11) with the bound on $RMSE(\boldsymbol{y}^{out})$ in Theorem 3.1, we observe that the additional terms induced by privacy are:

$$\frac{||M_{post}||_{2,2}}{\sqrt{T-T_0}}\frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1} + \frac{4T_0\sqrt{(8+n)n\sigma^2}}{\lambda\epsilon_1} + \frac{\sqrt{2n}\psi}{\epsilon_2} + \frac{4T_0\sqrt{2(8+n)}}{\lambda\epsilon_1\epsilon_2}. \qquad (12)$$

Then, using the fact that $\frac{||M_{post}||_{2,2}}{\sqrt{T-T_0}} \leq \sqrt{n}$ and setting $\epsilon := \epsilon_1 = \epsilon_2$ and $\lambda = O(T_0)$, Equation (12) can be bounded by,

$$\frac{4T_0\sqrt{(8+n)n}}{\lambda\epsilon} + \frac{4T_0\sqrt{(8+n)n\sigma^2}}{\lambda\epsilon} + \frac{\sqrt{2n}\psi}{\epsilon} + \frac{4T_0\sqrt{2(8+n)}}{\lambda\epsilon^2} = O\left(\frac{n}{\epsilon} + \frac{\sqrt{n}}{\epsilon^2}\right) = O\left(\frac{n}{\epsilon}\right) \text{ for } \epsilon \geq 1/\sqrt{n}.$$

Thus we conclude that the cost of privacy in the $DPSC_{out}$ algorithm is at most a factor of $O(\frac{1}{\epsilon})$. The restriction to $\epsilon \geq \frac{1}{\sqrt{n}}$ is consistent with standard practice in both theoretical and practical deployments of differential privacy, and thus is effectively without loss.

C.2 Closed-form bound on RMSE of Output Perturbation

In this section, we impose assumptions on the underlying data distribution to extend Theorem 3.2 to provide an explicit closed-form bound on the RMSE. Throughout this section, we make the following three mild assumptions of the distribution of X, which are required to achieve this closed-form expression:

Assumption C.2 *X_{pre} takes values in a k-dimensional subspace E for some small $k \ll \min\{n, T_0\}$.*

Assumption C.3 *The distribution of X_{pre} over E is isotropic, hence the covariance matrix $Cov(X_{pre}) = \Sigma = P_E$ where P_E is an orthogonal projection matrix onto E.*

Assumption C.4 *The distribution of $\boldsymbol{x_t} \in \mathbb{R}^n$ is supported in some centered Euclidean ball with radius $O(\sqrt{k})$.*

These assumptions are only slightly stronger than those commonly made in theory Amjad et al. (2018, 2019) and that typically hold in practice Udell and Townsend (2019). The first assumption means that X_{pre} is low rank. Assuming X_{pre} to be *approximately* low rank is a common practice in synthetic control literature Amjad et al. (2018, 2019). Indeed, most large matrices in practice are approximately low-rank Udell and Townsend (2019). Hence, we only further assume that it is *exactly* rank k for some small k. The second assumption allows us to apply useful mathematical properties: $||P_E||_2 = 1$ and $trace(P_E) = k$. Then, $\mathbb{E}[X_{pre}X_{pre}^\top] = trace(\Sigma) = k$ and, using Markov's inequality, we know that most of the distribution mass should be within a ball of radius $\sqrt{m}$ for $m = O(k)$. Hence, the third assumption asserts that not *most* but *all* the probability mass should lie within that ball, i.e., $||X||_{2,2} = O(\sqrt{k})$ almost surely.

Corollary C.5 provides a closed-form bound on the RMSE of $\boldsymbol{y}^{out}$ under these assumptions.

Corollary C.5 *If Assumptions C.2, C.3, and C.4 hold, then for all $\xi \in (0,1)$ and $t \geq 1$, with probability at least $1 - n^{-t^2}$, if $T_0 \geq C(t/\xi)^2 k \log n$, we have*

$$RMSE(\boldsymbol{y}^{out}) \leq \sqrt{n}\left(\frac{(\sqrt{2n\sigma^2} + \sqrt{2n\sigma^2 s^2}\,)T_0 + \frac{\lambda}{2T_0}}{(1-\xi)T_0 + \frac{\lambda}{2T_0}} + \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right) + \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}\right).$$

To derive Corollary C.5 from Theorem 3.2, we only need to derive and apply bounds on $||M_{post}||_{2,2}$ and $\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2]$. As we did before, we bound the first term using

$$||M_{post}||_{2,2} \leq ||M_{post}||_F \leq \sqrt{n(T - T_0)},$$

and thus $\frac{||M_{post}||_{2,2}}{\sqrt{T-T_0}} \leq \sqrt{n}$. Therefore, the key step is to bound $\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2]$. The following lemma provides the required bound on this term to prove Corollary C.5. The remainder of this section will be devoted to providing a proof sketch for Lemma C.6. A full proof is presented in Appendix F.3.

Lemma C.6 *Let $\boldsymbol{f}^{reg} = (X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}X_{pre}\boldsymbol{y}_{pre}$ be the Ridge regression coefficients and let $\boldsymbol{f}$ be the true coefficients. If Assumptions C.2, C.3, and C.4 hold, then for all $\xi \in (0,1)$ and $t \geq 1$, with probability at least $1 - n^{-t^2}$, if $T_0 \geq C(t/\xi)^2 k \log n$, we have,*

$$\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] \leq \frac{(\sqrt{2n\sigma^2} + \sqrt{2n\sigma^2 s^2})T_0 + \frac{\lambda}{2T_0}}{(1-\xi)T_0 + \frac{\lambda}{2T_0}}.$$

[Proof sketch of Lemma C.6.] First we can expand $\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2]$:

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] &= \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}] + \mathbb{E}[\boldsymbol{f}^{reg}] - \boldsymbol{f}||_2] \\
&\leq \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] + \mathbb{E}[||\mathbb{E}[\boldsymbol{f}^{reg}] - \boldsymbol{f}||_2] \\
&= \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] + \mathbb{E}[||Bias(\boldsymbol{f}^{reg})||_2]. \qquad (13)
\end{aligned}$$

We can bound these two terms separately as:

$$Bias(\boldsymbol{f}^{reg}) \leq \frac{\lambda}{2T_0}||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2} \quad \text{and}$$
$$\mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] \leq \mathbb{E}[||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2} \cdot ||X_{pre}\boldsymbol{z} - X_{pre}Z^\top \boldsymbol{f}||_2].$$

This can be combined back with Equation (13) to yield,

$$\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] \leq \mathbb{E}[||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2} \cdot (||X_{pre}\boldsymbol{z} - X_{pre}Z^\top \boldsymbol{f}||_2 + \frac{\lambda}{2T_0})]. \tag{14}$$

Next, we use our assumptions on the data distribution to prove the following lemma about $||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2}$.

Lemma C.7 *If Assumptions C.2, C.3, and C.4 hold, then for all $\xi \in (0,1)$ and $t \geq 1$, with probability at least $1 - n^{-t^2}$ and $T_0 \geq C(t/\xi)^2 k \log n$, we have*

$$||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2} \leq \frac{1}{(1-\xi)T_0 + \frac{\lambda}{2T_0}}.$$

To prove Lemma C.7, we use the following lemma about concentration of random matrices.

Lemma C.8 (Corollary 5.52 of Vershynin (2010)) *Consider a distribution in $\mathbb{R}^n$ with covariance matrix Σ, and supported in some centered Euclidean ball whose radius we denote $\sqrt{m}$. Let T_0 be the number of samples and define the sample covariance matrix $\Sigma_{T_0} = \frac{1}{T_0}XX^\top$. Let $\xi \in (0,1)$ and $t \geq 1$. Then with probability at least $1 - n^{-t^2}$, one has,*

$$\textit{If} \quad T_0 \geq C(t/\xi)^2 ||\Sigma||_{2,2}^{-1} m \log n \quad \textit{then } ||\Sigma_{T_0} - \Sigma||_{2,2} \leq \xi ||\Sigma||_{2,2},$$

where C is an absolute constant.

We instantiate Lemma C.8 using our assumptions that $||\Sigma||_{2,2} = ||P_E||_{2,2} = 1$ and the distribution is supported within some centered Euclidean ball with radius $\sqrt{O(k)}$ to get that with probability at least $1 - n^{-t^2}$ and $T_0 \geq C(t/\xi)^2 k \log n$,

$$||\frac{1}{T_0}X_{pre}X_{pre}^\top - \Sigma||_{2,2} \leq \xi.$$

We then use this to show that

$$||X_{pre}X_{pre}^\top - T_0 I||_{2,2} \leq \xi T_0,$$

and thus all eigenvalues of $(X_{pre}X_{pre}^\top - T_0 I)$ must be at most ξT_0, so

$$(1-\xi)T_0 \leq \boldsymbol{\lambda_{min}}(X_{pre}X_{pre}^\top) \leq (1+\xi)T_0.$$

Finally, we can complete the proof of Lemma C.7, by observing that,

$$||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2} = \frac{1}{|\boldsymbol{\lambda_{min}}(X^\top X) + \frac{\lambda}{2T_0}|} \leq \frac{1}{(1-\xi)T_0 + \frac{\lambda}{2T_0}}.$$

Returning to Equation (14), we can use this bound — along with the model properties specified in Equation (1) that each element of $\boldsymbol{z}$ and Z has mean 0, variance σ^2, and support $[-s, s]$ — to obtain the desired bound:

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] &\leq \frac{1}{(1-\xi)T_0 + \frac{\lambda}{2T_0}} \mathbb{E}[||X_{pre}\boldsymbol{z} - X_{pre}Z^\top \boldsymbol{f}||_2 + \frac{\lambda}{2T_0}] \\
&\leq \frac{1}{(1-\xi)T_0 + \frac{\lambda}{2T_0}} \left((\sqrt{nT_0} + \sqrt{nT_0 s^2}) \mathbb{E}[||\boldsymbol{z} - Z^\top \boldsymbol{f}||_2] + \frac{\lambda}{2T_0} \right) \\
&\leq \frac{1}{(1-\xi)T_0 + \frac{\lambda}{2T_0}} \left((\sqrt{nT_0} + \sqrt{nT_0 s^2}) \sqrt{2T_0 \sigma^2} + \frac{\lambda}{2T_0} \right) \\
&\leq \frac{(\sqrt{2n\sigma^2} + \sqrt{2n\sigma^2 s^2})T_0 + \frac{\lambda}{2T_0}}{(1-\xi)T_0 + \frac{\lambda}{2T_0}}.
\end{aligned}$$

D Privacy Guarantees of $DPSC_{obj}$

In this section, we prove Theorem 3.4, that $DPSC_{obj}$ is $(\epsilon_1+\epsilon_2, \delta)$-differentially private. This proof relies on composition of the (ϵ_1, δ)-DP learning step and the $(\epsilon_2, 0)$-DP prediction step. The prediction step is identical to that of Algorithm 2, so the privacy of this step follows immediately from Lemma 3.9 (that $\tilde{X}_{post}$ is computed in an $(\epsilon_2, 0)$-DP manner) and post-processing on the DP output of Step 1. All the remains to be shown is that $\boldsymbol{f}^{obj}$ is computed in an (ϵ_1, δ)-DP manner (Theorem D.1), and then Theorem 3.4 will follow by basic composition.

Theorem D.1 *Step 1 of Algorithm 3 that computes $\boldsymbol{f}^{obj}$ is (ϵ_1, δ)-differentially private.*

At a high-level, the privacy of $\boldsymbol{f}^{obj}$ comes from a carefully modified instantiation of the Objective Perturbation algorithms of Chaudhuri et al. (2011); Kifer et al. (2012), with novel sensitivity analysis, again due to the transposed regression setting of synthetic control (i.e., along columns not rows), where privacy is still required along the rows.

More formally, we start with the standard Ridge Regression objective function $J(\boldsymbol{f})$, that can be separated into the MSE loss function $\mathcal{L}(\boldsymbol{f})$ and the regularization term $r(\boldsymbol{f}) = \frac{\lambda}{2T_0}||\boldsymbol{f}||_2^2$ as follows:

$$J(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}) + r(\boldsymbol{f}) = \frac{1}{T_0}||y_{pre} - X_{pre}^\top \boldsymbol{f}||_2^2 + \frac{\lambda}{2T_0}||\boldsymbol{f}||_2^2.$$

The Objective Perturbation method modifies $J(\boldsymbol{f})$ by adding two terms: an additional regularization term and a noise term to ensure privacy:

$$J^{obj}(\boldsymbol{f}) = J(\boldsymbol{f}) + \frac{\Delta}{2T_0}||\boldsymbol{f}||_2^2 + \frac{1}{T_0}\boldsymbol{b}^\top \boldsymbol{f} = \mathcal{L}(\boldsymbol{f}) + \frac{\lambda+\Delta}{2T_0}||\boldsymbol{f}||_2^2 + \frac{1}{T_0}\boldsymbol{b}^\top \boldsymbol{f},$$

where $\boldsymbol{b}$ is a random vector drawn from a high-dimensional Laplace distribution if $\delta = 0$, and from a multivariate Gaussian distribution if $\delta > 0$.

Notice that $J^{obj}(\boldsymbol{f})$ is strongly convex (for any $\Delta \geq 0$) and differentiable. Hence, for any given input dataset $\mathcal{D} = (X_{pre}, y_{pre})$ and any fixed parameters $(\lambda, \epsilon_1, \epsilon_2, \delta)$, there exists a bijection between a realized value of the noise term $\boldsymbol{b}$ and $\boldsymbol{f}^{obj} := \arg\min_{\boldsymbol{f}} J^{obj}(\boldsymbol{f})$ given that realized $\boldsymbol{b}$.[2] We can then use this bijection to analyze the distribution over outputs on neighboring databases via the (explicitly given) noise distribution.

To observe this bijection concretely, let $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ be noise value that must have been realized when database $\mathcal{D}$ was input and $\boldsymbol{\alpha} = \arg\min_{\boldsymbol{f}} J^{obj}(\boldsymbol{f})$ was output. We can derive a closed-form expression for $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ by computing the gradient of $J^{obj}(\boldsymbol{f})$, which should be zero when evaluated at $\boldsymbol{f} = \boldsymbol{\alpha}$ since $\boldsymbol{\alpha}$ is defined to be the minimizer of $J^{obj}(\boldsymbol{f})$:

$$\nabla J^{obj}(\boldsymbol{f})\big|_{\boldsymbol{f}=\boldsymbol{\alpha}} = \nabla\mathcal{L}(\boldsymbol{\alpha}) + \nabla r(\boldsymbol{\alpha}) + \frac{\Delta}{T_0}\boldsymbol{\alpha} + \frac{\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})}{T_0} \stackrel{!}{=} 0.$$

Rearranging the equation yields

$$\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}) = -\left(T_0\nabla\mathcal{L}(\boldsymbol{\alpha}) + T_0\nabla r(\boldsymbol{\alpha}) + \Delta\boldsymbol{\alpha}\right).$$

Now, consider two arbitrary neighboring databases $\mathcal{D}$ and $\mathcal{D}'$ and an arbitrary output value $\boldsymbol{\alpha}$. Similar to Chaudhuri et al. (2011), we can use, e.g., Billingsley (1995) to express the ratio of the probabilities of outputting $\boldsymbol{\alpha}$ on neighboring $\mathcal{D}$ and $\mathcal{D}'$ as:[3]

$$\frac{\Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D})}{\Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}')} = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))} \frac{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))|}{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))|} := \Gamma(\boldsymbol{\alpha}) \cdot \Phi(\boldsymbol{\alpha}; \Delta),$$

where we define $\Gamma(\boldsymbol{\alpha}) := \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))}$ and $\Phi(\boldsymbol{\alpha}; \Delta) := \frac{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))|}{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))|}$. In the remainder of the proof, we will bound $\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_0}$ and $\Phi(\boldsymbol{\alpha}; \Delta) \leq e^{\epsilon_1 - \epsilon_0}$ so that the product is bounded by e^{ϵ_1}.

The parameter Δ serves a role to divide the ϵ_1 budget between these two terms, by distinguishing between two cases. In the first case, ϵ_1 is large enough that we can choose $\Delta = 0$ and still have some privacy budget (ϵ_0) remaining to bound $\Gamma(\boldsymbol{\alpha})$. In the other case, if ϵ_1 is too small to bound $\Phi(\boldsymbol{\alpha}; \Delta)$ with $\Delta = 0$, then we divide the privacy budget equally between bounding $\Gamma(\boldsymbol{\alpha})$ and $\Phi(\boldsymbol{\alpha}; \Delta)$, and find an appropriate value for $\Delta > 0$.

First, we will show $\Phi(\boldsymbol{\alpha}; \Delta)$ is upper bounded by $e^{\epsilon_1 - \epsilon_0}$.

[2] For a simple analogy, consider the one-dimensional Laplace Mechanism on query f and database x, which outputs $y = f(x) + Lap(\Delta f/\epsilon)$. Given f and x, there is a bijection between noise terms and outputs since the noise term must equal $y - f(x)$.

[3] with abuse of notation to let Pr denote pdf for simplicity of presentation.

Lemma D.2 *If $\Delta = 0$ and $\epsilon_0 = \epsilon_1 - \log(1 + \frac{2c}{\lambda} + \frac{c^2}{\lambda^2})$, or if $\Delta = \frac{c}{e^{\epsilon_1/4}-1} - \lambda$ and $\epsilon_0 = \frac{\epsilon_1}{2}$, then $\Phi(\boldsymbol{\alpha}; \Delta) \leq e^{\epsilon_1 - \epsilon_0}$.*

We start with Lemma D.3 (proved in Appendix G.1), which bounds $\Phi(\boldsymbol{\alpha}; \Delta)$ as a function of λ, c, and Δ.

Lemma D.3 *For any $\Delta \geq 0$, $\Phi(\boldsymbol{\alpha}; \Delta) = \frac{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))|}{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))|} \leq (1 + \frac{c}{\lambda + \Delta})^2$.*

Next, we use this result to prove our desired bound that $\Phi(\boldsymbol{\alpha}; \Delta) \leq e^{\epsilon_1 - \epsilon_0}$. We do this by considering two cases. First, when $\Delta = 0$, then $\Phi(\boldsymbol{\alpha}; \Delta = 0) \leq 1 + \frac{2c}{\lambda} + \frac{c^2}{\lambda^2} \leq e^{\epsilon_1 - \epsilon_0}$ by design, where the first inequality comes from Lemma D.3 and the second inequality is a rearrangement of our choice of $\epsilon_0 = \epsilon_1 - \log(1 + \frac{2c}{\lambda} + \frac{c^2}{\lambda^2})$ in this case. In the second case, $\Delta = \frac{c}{e^{\epsilon_1/4}-1} - \lambda$. Plugging this Δ value into the bound of Lemma D.3 gives $\Phi(\boldsymbol{\alpha}; \Delta) \leq e^{\epsilon_1/2} = e^{\epsilon_1 - \epsilon_0}$, where the second inequality come from our choice of $\epsilon_0 = \epsilon_1/2$. Hence, in both cases, $\Phi(\boldsymbol{\alpha}; \Delta) \leq e^{\epsilon_1 - \epsilon_0}$.

Then, we bound $\Gamma(\boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))}$. Note that this term depends only on the noise distribution, and not on the value of Δ. Algorithm 3 offers two options of noise distributions: Laplace noise when $\delta = 0$, and Gaussian noise when $\delta > 0$.

In the case of Laplace noise, the bound that $\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_0}$ follows immediately from the Laplace mechanism instantiated with privacy parameter ϵ_0 and Lemma B.4 to bound the sensitivity. The following lemma is proved in Appendix G.2.

Lemma D.4 *When $\boldsymbol{b}$ is sampled according to pdf $p(\boldsymbol{b}; \beta) \propto \exp\left(-\frac{||\boldsymbol{b}||_2}{\beta}\right)$, where $\beta = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$, then $\Gamma(\boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))} \leq e^{\epsilon_0}$.*

The two different β values come from two different upper bounds on the sensitivity, and the minimum value will give a tighter bound.

In the case where $\delta > 0$ and the Gaussian Mechanism is used, we cannot simply bound $\Gamma(\boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))}$ with probability 1. Instead, the bound must incorporate the δ term to bound $\Gamma(\boldsymbol{\alpha})$ with probability $1 - \delta$ over the internal randomness of the algorithm, as in Lemma D.5, formally proven in Appendix G.3.

Lemma D.5 *When $\boldsymbol{b} \sim \mathcal{N}(0, \beta^2 I_n)$, where $\beta = \frac{4T_0\sqrt{8+n}\sqrt{2\log\frac{2}{\delta} + \epsilon_0}}{\epsilon_0}$, then $\Gamma(\boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))} \leq e^{\epsilon_0}$ with probability at least $1 - \delta$.*

Finally, we combine the bounds on $\Phi(\boldsymbol{\alpha}; \Delta)$ and $\Gamma(\boldsymbol{\alpha})$ to complete the proof. When $\delta = 0$ with Laplace noise, Lemmas D.2 and D.4 combine immediately to give that $\Phi(\boldsymbol{\alpha}; \Delta)\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_1 - \epsilon_0 + \epsilon_0} = e^{\epsilon_1}$. When $\delta > 0$ and Gaussian noise is used, we define $\mathcal{G}$ to be the good event that $\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_0}$, which we know from Lemma D.5 will happen with at least probability $1 - \delta$. Then conditioned on $\mathcal{G}$ we have,

$$\frac{\Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}, \mathcal{G})}{\Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}', \mathcal{G})} = \Gamma(\boldsymbol{\alpha}) \cdot \Phi(\boldsymbol{\alpha}; \Delta) \leq e^{\epsilon_0} \cdot e^{\epsilon_1 - \epsilon_0} \leq e^{\epsilon_1}.$$

We can then use this fact to derive our desired (unconditioned) privacy bound:

$$\begin{aligned}
\Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}) &= \Pr(\mathcal{G}) \cdot \Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}, \mathcal{G}) + \Pr(\overline{\mathcal{G}}) \cdot \Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}, \overline{\mathcal{G}}) \\
&\leq e^{\epsilon_1} \Pr(\mathcal{G}) \cdot \Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}', \mathcal{G}) + \delta \\
&\leq e^{\epsilon_1} \Pr(\boldsymbol{f}^{obj} = \boldsymbol{\alpha} \mid \mathcal{D}') + \delta.
\end{aligned}$$

Hence, $\boldsymbol{f}^{obj}$ in Algorithm 3 is (ϵ_1, δ)-DP and the final output $\boldsymbol{y}^{obj}$ is $(\epsilon_1 + \epsilon_2, \delta)$-DP by composition.

E Accuracy Guarantees of $DPSC_{obj}$

In this section we analyze the accuracy of $DPSC_{obj}$. We first prove Theorem 3.5, restated below for convenience.

Theorem 3.5 *The estimator $\boldsymbol{y}^{obj}$ output by Algorithm 3 satisfies:*

$$\begin{aligned}
RMSE(\boldsymbol{y}^{obj}) \leq \frac{||M_{post}||_2}{\sqrt{T - T_0}} &\left(\mathbb{E}[||(\boldsymbol{f}^{reg} - \boldsymbol{f})||_2] + \frac{2}{\lambda + \Delta} \mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0} \left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta} \right) 2T_0^2 \sqrt{n} \right) \\
&+ \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2} \right) \left(\sqrt{n}\psi + \frac{2}{\lambda + \Delta} \mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0} \left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta} \right) 2T_0^2 \sqrt{n} \right),
\end{aligned}$$

where $||\boldsymbol{f}^{reg}||_\infty \leq \psi$ *for some* $\psi > 0$*, and* $\mathbb{E}[||\boldsymbol{b}||_2] = \sqrt{\frac{nT_0 4\sqrt{8+n}\sqrt{2\log\frac{2}{\delta}+\epsilon_0}}{\epsilon_0}}$ *for Gaussian noise (*$\delta > 0$ *case) and* $\mathbb{E}[||\boldsymbol{b}||_2] = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$ *for Laplace noise (*$\delta = 0$ *case), and* ϵ_0*, and* Δ *are computed internally by the algorithm.*

This theorem gives bounds on the predicted post-intervention target vector $\boldsymbol{y}^{obj}$, as measured by RMSE. Similar to Theorem 3.2, this result is stated in full generality with respect to the distribution of data and the latent variables, and thus the bound depends on terms such as $||M_{post}||_{2,2}$ and $\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2]$. Section E.1 provides proof of this result, with omitted detailed deferred to Appendix G.

Comparing the bound of Theorem 3.5 to that of Theorem 3.2 for output perturbation, we see that the difference comes only from the respective terms $\mathbb{E}[||(\boldsymbol{f}^{(out\vee obj)} - \boldsymbol{f}^{reg})||_2]$. For output perturbation, the error $\boldsymbol{f}^{out} - \boldsymbol{f}^{reg}$ is simply the noise directly added to the output, so the expected norm of the error is simply the expected norm of the noise, $a = \frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}$. For objective perturbation, the interpretation of these error terms is less straightforward and is instead bounded using Lemma E.1. As a simple case for comparison, when $\Delta = 0$ and $\delta = 0$ (i.e., using Laplace noise), the expected difference becomes $\mathbb{E}[||(\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2] \leq \min\{\frac{8T_0\sqrt{8+n}}{\lambda\epsilon_0}, \frac{2c\sqrt{n}+8T_0}{\lambda\epsilon_0}\}$. If the first term is the smaller of the two, then $\mathbb{E}[||(\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2]$ is bigger than $\mathbb{E}[||(\boldsymbol{f}^{out} - \boldsymbol{f}^{reg})||_2]$ since the denominator is smaller ($\epsilon_0 < \epsilon_1$, assuming the same ϵ_1 values for comparison) and the numerator is bigger due to the factor of 2. If the second term is the minimum, then the upper bound on error is hard to compare as both the denominator and the numerator are (asymptotically) bigger for output perturbation. In case of $\epsilon_0 = \frac{eps_1}{2}$, the expected difference of output perturbation becomes $O(T_0\sqrt{n})$ and that of objective perturbation becomes $O(T_0 + \sqrt{n})$. In this case, we may expect the objective perturbation to yield a better RMSE for a reasonably big T_0 and n.

E.1 Accuracy of post-intervention prediction via objective perturbation $\boldsymbol{y}^{obj}$

We will prove Theorem 3.5, which upper bounds the Root Mean Squared Error (RMSE) of $\boldsymbol{y}^{obj}$, defined as:

$$RMSE(\boldsymbol{y}^{obj}) = \frac{1}{\sqrt{T-T_0}}\mathbb{E}[||\boldsymbol{y}^{obj} - \boldsymbol{m_{post}}||_2].$$

Using the facts that $\boldsymbol{y}^{obj} = \tilde{X}_{post}^\top \boldsymbol{f}^{obj}$, $\tilde{X}_{post} = X_{post} + M_{post} + Z_{post}$, and $\boldsymbol{m}_{post} = M_{post}^\top \boldsymbol{f}$ (by Equation (2)), we can bound the expectation as follows:

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{y}^{obj} - \boldsymbol{m}_{post}||_2] &= \mathbb{E}[||\tilde{X}_{post}^\top \boldsymbol{f}^{obj} - M_{post}^\top \boldsymbol{f}||_2] \\
&= \mathbb{E}[||\tilde{X}_{post}^\top \boldsymbol{f}^{obj} - \tilde{X}_{post}^\top \boldsymbol{f}^{reg} + \tilde{X}_{post}^\top \boldsymbol{f}^{reg} - M_{post}^\top \boldsymbol{f}||_2] \\
&= \mathbb{E}[||(M_{post} + Z_{post} + W_{post})^\top (\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg}) + (M_{post} + Z_{post} + W_{post})^\top \boldsymbol{f}^{reg} - M_{post}^\top \boldsymbol{f}||_2] \\
&\leq \mathbb{E}[||(M_{post} + Z_{post} + W_{post})^\top (\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2] \\
&\quad + \mathbb{E}[||M_{post}^\top (\boldsymbol{f}^{reg} - \boldsymbol{f})||_2] + \mathbb{E}[||(Z_{post} + W_{post})^\top \boldsymbol{f}^{reg}||_2], \qquad (15)
\end{aligned}$$

where the first equality is due to the definition of $\boldsymbol{y}^{obj}$, the second equality adds and subtracts the same term, the third equality collects terms and plugs in the expression for $\tilde{X}_{post}$, and the final step is due to triangle inequality.

Lemma C.1 already bounds the last two terms because they do not involve $\boldsymbol{f}^{obj}$ and Step 2 of Algorithms 2 and 3 are the same. Specifically, we know that,

$$\mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg} - \boldsymbol{f})||_2] \leq ||M_{post}||_{2,2}\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] \text{ and } \mathbb{E}[||(Z_{post}^\top + W_{post}^\top)\boldsymbol{f}^{reg}||_2] \leq \sqrt{n}\psi(\sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2}). \qquad (16)$$

Thus we only need to bound the first term:

$$\begin{aligned}
&\mathbb{E}[||(M_{post} + Z_{post} + W_{post})^\top (\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2] \\
&\quad \leq \mathbb{E}[||M_{post}^\top (\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2] + \mathbb{E}[||(Z_{post} + W_{post})^\top (\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2] \\
&\quad \leq ||M_{post}||_2 \mathbb{E}[||\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg}||_2] + \mathbb{E}[||Z_{post} + W_{post}||_2]\mathbb{E}[||(\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2] \\
&\quad \leq ||M_{post}||_2 \mathbb{E}[||\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg}||_2] + (\sqrt{n(T-T_0)\sigma^2} + \frac{2\sqrt{T-T_0}}{\epsilon_2})\mathbb{E}[||(\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2], \qquad (17)
\end{aligned}$$

where the first step is simply triangle inequality, the second step is due to the independence of Z, W and $\boldsymbol{f}^{obj}$, and the third step comes from the proof of Lemma C.1 (see Appendix F.2), where $\mathbb{E}[||Z_{post} + W_{post}||_2]$ was bounded as an intermediate step.

Thus we only need to derive a bound on $\mathbb{E}[||\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg}||_2]$, which we do in Lemma E.1 (formally proven in Appendix G.4) to complete the proof.

Lemma E.1 *The ℓ_2 distance between $\boldsymbol{f}^{obj}$ and $\boldsymbol{f}^{reg}$ satisfies:*

$$\mathbb{E}[||\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg}||_2] \leq \frac{2}{\lambda + \Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta}\right) 2T_0^2\sqrt{n},$$

where $\boldsymbol{b}$ and Δ are computed internally by Algorithm 3.

Combining Equations (15), (16), and (17) with Lemma E.1 completes the proof of Theorem 3.5:

$$\begin{aligned}
RMSE(\boldsymbol{y}^{obj}) &\leq \frac{||M_{post}||_2}{\sqrt{T - T_0}}\left(\mathbb{E}[||(\boldsymbol{f}^{reg} - \boldsymbol{f})||_2] + \mathbb{E}[||(\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2]\right) \\
&\quad + \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \mathbb{E}[||(\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg})||_2]\right) \\
&\leq \frac{||M_{post}||_2}{\sqrt{T - T_0}}\left(\mathbb{E}[||(\boldsymbol{f}^{reg} - \boldsymbol{f})||_2] + \frac{2}{\lambda + \Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta}\right) 2T_0^2\sqrt{n}\right) \\
&\quad + \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \frac{2}{\lambda + \Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta}\right) 2T_0^2\sqrt{n}\right),
\end{aligned}$$

where $\mathbb{E}[||\boldsymbol{b}||_2] = \sqrt{n\beta} = \sqrt{\frac{nT_0\zeta\sqrt{2\log\frac{2}{\delta} + \epsilon_0}}{\epsilon_0}}$ for Gaussian noise and $\mathbb{E}[||\boldsymbol{b}||_2] = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$ for Laplace noise.

E.2 Closed-form bound on RMSE of Objective Perturbation

Using similar analysis as in Section C.2, we can extend Theorem 3.5 to obtain the following closed-form accuracy bound that depends only on explicit input parameters, under the same distributional assumptions.

Corollary E.2 *If Assumptions C.2, C.3, and C.4 hold, then for all $\xi \in (0, 1)$ and $t \geq 1$, with probability at least $1 - n^{-t^2}$, if $T_0 \geq C(t/\xi)^2 k \log n$, we have*

$$\begin{aligned}
RMSE(\boldsymbol{y}^{out}) \leq \sqrt{n}\Bigg(&\frac{(\sqrt{2n\sigma^2} + \sqrt{2n\sigma^2 s^2})T_0 + \frac{\lambda}{2T_0}}{(1 - \xi)T_0 + \frac{\lambda}{2T_0}} + \frac{2}{\lambda + \Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta}\right) 2T_0^2\sqrt{n}\Bigg) \\
&+ \left(\sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2}\right)\left(\sqrt{n}\psi + \frac{2}{\lambda + \Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta}\right) 2T_0^2\sqrt{n}.\right),
\end{aligned}$$

where $||\boldsymbol{f}^{reg}||_\infty \leq \psi$ for some $\psi > 0$, and $\mathbb{E}[||\boldsymbol{b}||_2] = \sqrt{\frac{nT_0 4\sqrt{8+n}\sqrt{2\log\frac{2}{\delta} + \epsilon_0}}{\epsilon_0}}$ for Gaussian noise ($\delta > 0$ case) and $\mathbb{E}[||\boldsymbol{b}||_2] = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$ for Laplace noise ($\delta = 0$ case), and ϵ_0 and Δ are computed internally by the algorithm.

The additional terms that arise due to the noise required to guarantee differential privacy in this setting, relative to the bound on $RMSE(\boldsymbol{y}^{reg})$ in Equation (11), are:

$$(\sqrt{n} + \sqrt{n\sigma^2} + \frac{\sqrt{2}}{\epsilon_2})\left(\frac{2}{\lambda + \Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta \neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \Delta}\right) 2T_0^2\sqrt{n}\right) + \frac{\sqrt{2n}}{\epsilon_2}\psi. \tag{18}$$

To analyze this expression in a simplified way, assume the regularization parameter is $\lambda = O(T_0)$ so $\frac{T_0}{\lambda} = O(1)$, and that Laplace noise was used (i.e., $\delta = 0$), so that $\mathbb{E}[||\boldsymbol{b}||_2] = O(\frac{T_0\sqrt{n}}{\epsilon_0})$. Then the first parenthesis of Equation (18) becomes

$O(\sqrt{n}+\frac{1}{\epsilon_2})$, the second parenthesis becomes $O(\frac{T_0\sqrt{n}}{\epsilon_0}+T_0\sqrt{n})$, and the additive term becomes $O(\frac{\sqrt{n}}{\epsilon_2})$. Since $\epsilon_0 < \epsilon_1$, we replace ϵ_0 by ϵ_1 in the bounds. Then Equation (18) is $O(\frac{T_0 n}{\epsilon_1}+\frac{T_0\sqrt{n}}{\epsilon_1\epsilon_2})$ from the product of two parentheses, and omitting the additive term, which is asymptotically dominated by the others.

Comparing to the cost of privacy in Output Perturbation in Corollary C.5, we see that the bound in Corollary C.5 does not depend on T_0. This additional dependence on T_0 arises for Objective Perturbation from the second parenthesis containing $\mathbb{E}[||\boldsymbol{b}||_2]$ and the indicator function, which is absent in the output perturbation case.

F Omitted Proofs for $DPSC_{out}$

F.1 Proof of Lemma B.4

Lemma B.4 *Let $g(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}, \mathcal{D}') - \mathcal{L}(\boldsymbol{f}, \mathcal{D})$ for two arbitrarily neighboring databases $\mathcal{D}, \mathcal{D}'$. Then,*

$$\max_{\boldsymbol{f}} \|\nabla g(\boldsymbol{f})\| \leq 4\sqrt{8+n}.$$

We first re-arrange $g(\boldsymbol{f})$ in a way that makes it easier to compute the gradient. Let i be the index of the record that differs between $\mathcal{D}$ and $\mathcal{D}'$.

$$\begin{aligned}
g(\boldsymbol{f}) &= \mathcal{L}(\boldsymbol{f}, \mathcal{D}') - \mathcal{L}(\boldsymbol{f}, \mathcal{D}) \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[\left(\sum_{k=1}^{n} {x'}_{k,t} f_k\right) - y_t\right]^2 - \frac{1}{T_0}\sum_{t=1}^{T_0}\left[\left(\sum_{k=1}^{n} x_{k,t} f_k\right) - y_t\right]^2 \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[\left(\sum_{j\neq i} {x'}_{j,t} f_j\right) - y_t + {x'}_{i,t} f_i\right]^2 - \frac{1}{T_0}\sum_{t=1}^{T_0}\left\{\left(\sum_{j\neq i} x_{j,t} f_j\right) - y_t + x_{i,t} f_i\right\}^2 \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[2\left(\sum_{j\neq i} x_{j,t} f_j - y_t\right)({x'}_{i,t} - x_{i,t}) f_i + ({x'}_{i,t}^2 - x_{i,t}^2) f_i^2\right] \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[2\left(\sum_{j\neq i} x_{j,t} f_j - y_t\right) + ({x'}_{i,t} + x_{i,t}) f_i\right]({x'}_{i,t} - x_{i,t})\, f_i \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[\left(\boldsymbol{x_t}^\top \boldsymbol{f} - y_t\right) + \left(\boldsymbol{x_t}^\top \boldsymbol{f} - x_{i,t} f_i + {x'}_{i,t} f_i - y_t\right)\right]({x'}_{i,t} - x_{i,t}) f_i
\end{aligned} \tag{19}$$

The second equality comes from the definition of the Ridge regression loss function $\mathcal{L}$; in the third step we pull out the record i that differs between $\mathcal{D}$ and $\mathcal{D}'$; the fourth step combines the sums and cancels terms, including the observation that $\sum_{j\neq i} x_{j,t} f_j = \sum_{j\neq i} x'_{j,t} f_j$. The final two steps also involve rearranging terms.

For notational ease, we define two additional terms,

$$D_t := \boldsymbol{x_t}^\top \boldsymbol{f} - y_t \quad \text{and} \quad E_t := (x'_{i,t} - x_{i,t}) f_i.$$

Then, Equation (19) becomes

$$g(\boldsymbol{f}) = \frac{1}{T_0}\sum_{t=1}^{T_0}(2D_t + E_t)E_t.$$

We will take the partial derivatives of D_t and E_t with respect to both f_i (the index of the data entry that differs between $\mathcal{D}$ and $\mathcal{D}'$) and f_j for $j \neq i$, and then combine these to arrive at the gradient of $g(\boldsymbol{f})$:

$$\frac{\partial D_t}{\partial f_i} = x_{i,t}; \quad \frac{\partial D_t}{\partial f_j} = x_{j,t}; \quad \frac{\partial E_t}{\partial f_i} = x'_{i,t} - x_{i,t}; \quad \frac{\partial E_t}{\partial f_j} = 0.$$

Now, we compute the derivative of $g(\boldsymbol{f})$ with respect to f_i.

$$\begin{aligned}
\frac{\partial g(\boldsymbol{f})}{\partial f_i} &= \frac{1}{T_0}\sum_{t=1}^{T_0}\left\{\left(2\frac{\partial D_t}{\partial f_i}+\frac{\partial E_t}{\partial f_i}\right)E_t + (2D_t+E_t)\frac{\partial E_t}{\partial f_i}\right\} \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left\{(x'_{i,t}+x_{i,t})E_t + (2D_t+E_t)(x'_{i,t}-x_{i,t})\right\} \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left\{2x'_{i,t}(x'_{i,t}-x_{i,t})f_i + 2(\boldsymbol{x_t}^\top\boldsymbol{f}-y_t)(x'_{i,t}-x_{i,t})\right\} \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}2\left(x'_{i,t}f_i+\boldsymbol{x_t}^\top\boldsymbol{f}-y_t\right)(x'_{i,t}-x_{i,t}) \qquad (20)
\end{aligned}$$

Next, we compute the derivative of $g(\boldsymbol{f})$ with respect to f_j where j is the index of unchanged donors ($j \neq i$). There are fewer term in this derivative because $\frac{\partial E_t}{\partial f_j} = 0$.

$$\begin{aligned}
\frac{\partial g(\boldsymbol{f})}{\partial f_j} &= \frac{1}{T_0}\sum_{t=1}^{T_0}\left\{(2x_{j,t})\,E_t\right\} \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}2x_{j,t}(x'_{i,t}-x_{i,t})f_i \qquad (21)
\end{aligned}$$

Finally, we can use (20) and (21) to derive an upper bound for $||\nabla g(\boldsymbol{f})||_2$.

$$\begin{aligned}
||\nabla g(\boldsymbol{f})||_2^2 &= \left(\frac{\partial g(\boldsymbol{f})}{\partial f_i}\right)^2 + \sum_{j\neq i}\left(\frac{\partial g(\boldsymbol{f})}{\partial f_j}\right)^2 \\
&= \left(\frac{1}{T_0}\sum_{t=1}^{T_0}2\left(x'_{i,t}f_i+\boldsymbol{x_t}^\top\boldsymbol{f}-y_t\right)(x'_{i,t}-x_{i,t})\right)^2 + \sum_{j\neq i}\left(\frac{1}{T_0}\sum_{t=1}^{T_0}2x_{j,t}(x'_{i,t}-x_{i,t})f_i\right)^2 \\
&\leq \frac{1}{T_0}\sum_{t=1}^{T_0}\left[2(x'_{i,t}f_i+\boldsymbol{x_t}^\top\boldsymbol{f}-y_t)(x'_{i,t}-x_{i,t})\right]^2 + \sum_{j\neq i}\left[\frac{1}{T_0}\sum_{t=1}^{T_0}\left[2x_{j,t}(x'_{i,t}-x_{i,t})f_i\right]^2\right] \\
&= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[4(x'_{i,t}f_i+\boldsymbol{x_t}^\top\boldsymbol{f}-y_t)^2(x'_{i,t}-x_{i,t})^2 + \sum_{j\neq i}4x_{j,t}^2(x'_{i,t}-x_{i,t})^2f_i^2\right] \\
&= \frac{4}{T_0}\sum_{t=1}^{T_0}(x'_{i,t}-x_{i,t})^2\left[(x'_{i,t}f_i+\boldsymbol{x_t}^\top\boldsymbol{f}-y_t)^2 + \sum_{j\neq i}x_{j,t}^2f_i^2\right] \\
&= \frac{4}{T_0}\sum_{t=1}^{T_0}(x'_{i,t}-x_{i,t})^2\left[(\boldsymbol{x_t}^\top\boldsymbol{f}-y_t)^2 + 2x'_{i,t}f_i(\boldsymbol{x_t}^\top\boldsymbol{f}-y_t) + \sum_{k=1}^{n}x'^2_{k,t}f_i^2\right] \qquad (22)
\end{aligned}$$

The second equality comes from plugging in the partial derivatives computed in (20) and (21), the following inequality comes from applying Jensen's inequality, and the final three steps come from rearranging, expanding, and simplifying terms.

We can proceed by bounding the individual terms in (22) using the our modeling assumptions of Equation (3), which give us that:

$$(x'_{i,t}-x_{i,t})^2 \leq 4, \quad \text{and} \quad (\boldsymbol{x_t}^\top\boldsymbol{f}-y_t)^2 \leq 4, \quad \text{and} \quad 2x'_{i,t}f_i(\boldsymbol{x_t}^\top\boldsymbol{f}-y_t) \leq 4, \quad \text{and} \quad \sum_{k=1}^{n}x'^2_{k,t}f_i^2 \leq n.$$

Then $||\nabla g(\boldsymbol{f})||_2^2 \leq 128 + 16n$ and $||\nabla g(\boldsymbol{f})||_2 \leq 4\sqrt{8+n}$.

F.2 Proof of Lemma C.1

Lemma C.1 *The three terms in Equation* (8) *can be bounded as follows:*

$$\mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg}-\boldsymbol{f})||_2] \le ||M_{post}||_{2,2}\cdot\mathbb{E}[||\boldsymbol{f}^{reg}-\boldsymbol{f}||_2],$$

$$\mathbb{E}[||(Z^\top+W^\top)\boldsymbol{f}^{reg}||_2] \le \sqrt{n}\psi\left(\sqrt{n(T-T_0)\sigma^2}+\frac{2\sqrt{T-T_0}}{\epsilon_2}\right), \text{ and}$$

$$\mathbb{E}[||(M_{post}^\top+Z^\top+W^\top)\boldsymbol{v}||_2] \le \left(||M_{post}||_{2,2}+\sqrt{n(T-T_0)\sigma^2}+\frac{2\sqrt{T-T_0}}{\epsilon_2}\right)\frac{4T_0\sqrt{8+n}}{\lambda\epsilon_1}.$$

We prove these three bounds separately. Most steps follow from the sub-multiplicative norm property of Equation (9) and the bounds on the noise terms of Equation (10).

First,

$$\begin{aligned}\mathbb{E}[||M_{post}^\top(\boldsymbol{f}^{reg}-\boldsymbol{f})||_2] &\le \mathbb{E}[||M_{post}||_{2,2}\cdot||\boldsymbol{f}^{reg}-\boldsymbol{f}||_2]\\ &\le ||M_{post}||_{2,2}\mathbb{E}[\boldsymbol{f}^{reg}-\boldsymbol{f}||_2].\end{aligned}$$

Next,

$$\begin{aligned}\mathbb{E}[||(Z_{post}^\top+W^\top)\boldsymbol{f}^{reg}||_2] &\le \mathbb{E}[||Z_{post}+W||_2\cdot||\boldsymbol{f}^{reg}||_2]\\ &\le \mathbb{E}[||Z_{post}+W||_2\cdot\sqrt{n}||\boldsymbol{f}^{reg}||_\infty]\\ &\le \mathbb{E}[||Z_{post}+W||_2\cdot\sqrt{n}\psi]\\ &\le \sqrt{n}\psi\cdot\mathbb{E}[||Z_{post}||_2+||W||_2]\\ &\le \sqrt{n}\psi\cdot\mathbb{E}[||Z_{post}||_F+||W||_F]\\ &\le \sqrt{n}\psi\left(\sqrt{n(T-T_0)\sigma^2}+b\right)\\ &\le \sqrt{n}\psi\left(\sqrt{n(T-T_0)\sigma^2}+\frac{2\sqrt{T-T_0}}{\epsilon_2}\right),\end{aligned}$$

where the second step comes from the relationship between the ℓ_2 norm and the ℓ_∞ norm, and the third step comes from our definition that $||\boldsymbol{f}^{reg}||_\infty \le \psi$ for some $\psi > 0$.

Finally,

$$\begin{aligned}\mathbb{E}[||(M_{post}^\top+Z_{post}^\top+W^\top)\boldsymbol{v}||_2] &\le \mathbb{E}[||M_{post}+Z_{post}+W||_{2,2}||\boldsymbol{v}||_2]\\ &= \mathbb{E}[||M_{post}+Z_{post}+W||_{2,2}]\cdot\mathbb{E}[||\boldsymbol{v}||_2]\\ &\le \mathbb{E}[||M_{post}||_{2,2}+||Z_{post}||_{2,2}+||W||_{2,2}]\cdot\frac{4\sqrt{8+n}}{\lambda\epsilon_1}\\ &\le (||M_{post}||_{2,2}+\mathbb{E}[||Z_{post}||_F+||W||_F])\cdot\frac{4\sqrt{8+n}}{\lambda\epsilon_1}\\ &\le \left(||M_{post}||_{2,2}+\sqrt{n(T-T_0)\sigma^2}+\frac{2\sqrt{T-T_0}}{\epsilon_2}\right)\frac{4\sqrt{8+n}}{\lambda\epsilon_1},\end{aligned}$$

where the second step holds because Z_{post}, W and $\boldsymbol{v}$ are all independent of each other.

F.3 Proof of Lemma C.6

Lemma C.6 *Let $\boldsymbol{f}^{reg} = (X_{pre}X_{pre}^\top+\frac{\lambda}{2T_0}I)^{-1}X_{pre}\boldsymbol{y}_{pre}$ be the Ridge regression coefficients and let $\boldsymbol{f}$ be the true coefficients. If Assumptions C.2, C.3, and C.4 hold, then for all $\xi\in(0,1)$ and $t\ge 1$, with probability at least $1-n^{-t^2}$, if $T_0 \ge C(t/\xi)^2 k\log n$, we have,*

$$\mathbb{E}[||\boldsymbol{f}^{reg}-\boldsymbol{f}||_2] \le \frac{(\sqrt{2n\sigma^2}+\sqrt{2n\sigma^2 s^2})T_0+\frac{\lambda}{2T_0}}{(1-\xi)T_0+\frac{\lambda}{2T_0}}.$$

First we can expand $\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2]$:

$$\begin{aligned}\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] &= \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}] + \mathbb{E}[\boldsymbol{f}^{reg}] - \boldsymbol{f}||_2] \\ &\leq \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] + \mathbb{E}[||\mathbb{E}[\boldsymbol{f}^{reg}] - \boldsymbol{f}||_2] \\ &= \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] + \mathbb{E}[||Bias(\boldsymbol{f}^{reg})||_2], \end{aligned} \tag{23}$$

where

$$Bias(\boldsymbol{f}^{reg}) = \mathbb{E}[\boldsymbol{f}^{reg}] - \boldsymbol{f} = -\lambda(X_{pre}X_{pre}^\top + \lambda I)^{-1}\boldsymbol{f}.$$

Hence, we only need to bound the two terms: $||Bias(\boldsymbol{f}^{reg})||_2$ and $\mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2]$, which we do next. First,

$$\begin{aligned}||Bias(\boldsymbol{f}^{reg})||_2 &= ||-\lambda(X_{pre}X_{pre}^\top + \lambda I)^{-1}\boldsymbol{f}||_2 \\ &\leq \lambda||\boldsymbol{f}||_2||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2} \\ &\leq \lambda||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2},\end{aligned}$$

where the last inequality uses the fact that the ℓ_1 norm of $\boldsymbol{f}$ is 1, which also upper bound the ℓ_2 norm. Next,

$$\begin{aligned}\mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] &= \mathbb{E}[||\boldsymbol{f}^{reg} - (\boldsymbol{f} + Bias(\boldsymbol{f}^{reg}))||_2] \\ &= \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}X_{pre}\boldsymbol{y_{pre}} - \boldsymbol{f} + \lambda(X_{pre}X_{pre}^\top + \lambda I)^{-1}\boldsymbol{f}||_2] \\ &= \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}X_{pre}(M_{pre}^\top\boldsymbol{f} + \boldsymbol{z_{pre}}) - \boldsymbol{f} + \lambda(X_{pre}X_{pre}^\top + \lambda I)^{-1}\boldsymbol{f}||_2] \\ &= \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}X_{pre}(X_{pre}^\top\boldsymbol{f} - Z_{pre}^\top\boldsymbol{f} + \boldsymbol{z_{pre}}) - \boldsymbol{f} + \lambda(X_{pre}X_{pre}^\top + \lambda I)^{-1}\boldsymbol{f}||_2] \\ &= \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}(X_{pre}X_{pre}^\top + \lambda I)\boldsymbol{f} - \boldsymbol{f} \\ &\quad + (X_{pre}X_{pre}^\top + \lambda I)^{-1}(X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top\boldsymbol{f})||_2] \\ &= \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}(X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top\boldsymbol{f})||_2] \\ &\leq \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2} \cdot ||X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top\boldsymbol{f}||_2],\end{aligned}$$

where the first four steps come respectively from plugging in expressions of $\mathbb{E}[\boldsymbol{f}^{reg}]$, ($\boldsymbol{f}^{reg}$ and $Bias(\boldsymbol{f}^{reg})$), $\boldsymbol{y_{pre}}$, and M_{pre}. The fifth and sixth steps come from rearranging and canceling terms, and the final inequality comes from the submultiplicative norm property of Equation (9).

Plugging everything back to (23) yields,

$$\begin{aligned}\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] &\leq \mathbb{E}[||\boldsymbol{f}^{reg} - \mathbb{E}[\boldsymbol{f}^{reg}]||_2] + \mathbb{E}[||Bias(\boldsymbol{f}^{reg})||_2] \\ &\leq \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2} \cdot ||X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top\boldsymbol{f}||_2 + \lambda||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2}] \\ &= \mathbb{E}[||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2} \cdot (||X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top\boldsymbol{f}||_2 + \lambda)]\end{aligned} \tag{24}$$

Next, we use our assumptions on the data distribution to prove the following lemma about $||(X_{pre}X_{pre}^\top + \lambda I)^{-1}||_{2,2}$.

Lemma C.7 *If Assumptions C.2, C.3, and C.4 hold, then for all $\xi \in (0,1)$ and $t \geq 1$, with probability at least $1 - n^{-t^2}$ and $T_0 \geq C(t/\xi)^2 k \log n$, we have*

$$||(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0}I)^{-1}||_{2,2} \leq \frac{1}{(1-\xi)T_0 + \frac{\lambda}{2T_0}}.$$

[Proof of Lemma C.7] A key component of the proof of Lemma C.7 is the following lemma about concentration of random matrices.

Lemma C.8 (Corollary 5.52 of Vershynin (2010)) *Consider a distribution in $\mathbb{R}^n$ with covariance matrix Σ, and supported in some centered Euclidean ball whose radius we denote $\sqrt{m}$. Let T_0 be the number of samples and define the sample covariance matrix $\Sigma_{T_0} = \frac{1}{T_0}XX^\top$. Let $\xi \in (0,1)$ and $t \geq 1$. Then with probability at least $1 - n^{-t^2}$, one has,*

$$\textit{If} \quad T_0 \geq C(t/\xi)^2||\Sigma||_{2,2}^{-1}m\log n \quad \textit{then } ||\Sigma_{T_0} - \Sigma||_{2,2} \leq \xi||\Sigma||_{2,2},$$

where C is an absolute constant.

To instantiate Lemma C.8, we view the data X_{pre} as T_0 samples corresponding to the columns $\boldsymbol{x_t} \in \mathbb{R}^n, \ \forall t \in \{1, 2, \cdots, T_0\}$. We use our assumptions that X takes values in a k-dimensional subspace E, and $\Sigma = P_E$ where P_E is the orthogonal projection from $\mathbb{R}^n$ onto E. Then, the *effective rank* of Σ is $r(\Sigma) = \frac{trace(\Sigma)}{||\Sigma||_2} = k$ by definition, because $||\Sigma||_{2,2} = ||P_E||_{2,2} = \boldsymbol{\sigma}_{max}(P_E) = \boldsymbol{\lambda}_{max}(P_E) = 1$, since eigenvalues of an orthogonal projection matrix are either 0 or 1 as shown in Lemma 19 of Amjad et al. (2018). Then, $\mathbb{E}[||X||_{2,2}^2] = trace(\Sigma) = k||\Sigma||_{2,2} = k||P_E||_{2,2} = k$. Using Markov's inequality, most of the distribution should be within a ball of radius $\sqrt{m}$ where $m = O(k)$. Finally, let us assume that all the probability mass is within that ball, i.e., $||X||_{2,2} = O(\sqrt{k})$ almost surely. Then, Lemma C.8 holds with $T_0 \geq C(t/\epsilon)^2 k \log n$ samples. This is also noted in Remark 5.53 of Vershynin (2010).

To translate this to our setting, we see that with probability at least $1 - n^{-t^2}$, if $T_0 \geq C(t/\xi)^2 k \log n$, then

$$||\frac{1}{T_0} X_{pre} X_{pre}^\top - \Sigma||_{2,2} \leq \xi. \tag{25}$$

Since $\Sigma = P_E$ is an orthogonal projection matrix, $||P_E||_{2,2} = 1$. We apply triangle inequality to obtain,

$$||\frac{1}{T_0} X_{pre} X_{pre}^\top - P_E||_{2,2} \geq \left| ||\frac{1}{T_0} X_{pre} X_{pre}^\top||_{2,2} - ||P_E||_{2,2} \right| = \left| ||\frac{1}{T_0} X_{pre} X_{pre}^\top||_{2,2} - 1 \right| \geq ||\frac{1}{T_0} X_{pre} X_{pre}^\top - I||_{2,2}.$$

Combining this with Equation (25), we can bound

$$||\frac{1}{T_0} X_{pre} X_{pre}^\top - I||_{2,2} \leq \xi, \quad \text{or equivalently,} \quad ||X_{pre} X_{pre}^\top - T_0 I||_{2,2} \leq \xi T_0. \tag{26}$$

We will use this latter expression to obtain a lower bound on the minimum singular value of $X_{pre} X_{pre}^\top$, and then use it to bound $||(X_{pre} X_{pre}^\top + \lambda I)^{-1}||_{2,2}$ from above.

Note that since $||A||_{2,2}$ is the maximum singular value of matrix A, the upper bound of ξT_0 of Equation (26) should hold for all singular values of A. For symmetric matrices such as $X_{pre} X_{pre}^\top + T_0 I$, the singular values are also the absolute values of its eigenvalues. This means that all eigenvalues $\boldsymbol{\lambda_\star}$ of $X_{pre} X_{pre}^\top - T_0 I$ must satisfy $|\boldsymbol{\lambda_\star}(X_{pre} X_{pre}^\top - T_0 I)| \leq \xi T_0$. Therefore, this bound must also hold for the smallest eigenvalue $\boldsymbol{\lambda_{min}}(\cdot)$:

$$\begin{aligned}
& |\boldsymbol{\lambda_{min}}(X_{pre} X_{pre}^\top - T_0 I)| \leq \xi T_0 \\
\iff \quad & |\boldsymbol{\lambda_{min}}(X_{pre} X_{pre}^\top) - T_0| \leq \xi T_0 \\
\iff \quad & (1 - \xi) T_0 \leq \boldsymbol{\lambda_{min}}(X_{pre} X_{pre}^\top) \leq (1 + \xi) T_0
\end{aligned}$$

By plugging in the lower bound on the minimum singular value of $X_{pre} X_{pre}^\top$, we arrive at the desired bound to complete the proof of Lemma C.7.

$$\begin{aligned}
||(X_{pre} X_{pre}^\top + \lambda I)^{-1}||_2 &= \boldsymbol{\sigma_{max}}((X_{pre} X_{pre}^\top + \lambda I)^{-1}) \\
&= \frac{1}{\boldsymbol{\sigma_{min}}(X_{pre} X_{pre}^\top + \lambda I)} \\
&= \frac{1}{|\boldsymbol{\lambda_{min}}(X^\top X) + \lambda|} \\
&\leq \frac{1}{(1 - \xi) T_0 + \lambda}.
\end{aligned}$$

Returning to Equation (24), we can use this bound to obtain,

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] &\leq \mathbb{E}[||(X_{pre} X_{pre}^\top + \lambda I)^{-1}||_{2,2} \cdot (||X_{pre} \boldsymbol{z} - X_{pre} Z^\top \boldsymbol{f}||_2 + \lambda)] \\
&\leq \frac{1}{(1 - \xi) T_0 + \lambda} \mathbb{E}[||X_{pre} \boldsymbol{z} - X_{pre} Z^\top \boldsymbol{f}||_2 + \lambda]
\end{aligned} \tag{27}$$

The expectation term in Equation (27) becomes,

$$\begin{aligned}
\mathbb{E}[||X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top \boldsymbol{f}||_2 + \lambda] &= \mathbb{E}[||(M_{pre} + Z_{pre})\boldsymbol{z_{pre}} - (M + Z_{pre})Z_{pre}^\top \boldsymbol{f}||_2 + \lambda] \\
&\leq \mathbb{E}[||M_{pre}(\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f})||_2 + ||Z_{pre}(\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f})||_2 + \lambda] \\
&\leq ||M_{pre}||_{2,2}\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \mathbb{E}[||Z_{pre}||_{2,2} \cdot ||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \lambda \\
&= ||M_{pre}||_F\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \mathbb{E}[||Z_{pre}||_F \cdot ||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \lambda \\
&\leq \sqrt{nT_0}\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \sqrt{nT_0 s^2}\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \lambda \\
&= (\sqrt{nT_0} + \sqrt{nT_0 s^2})\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] + \lambda,
\end{aligned}$$

where the first step is plugging in for X_{pre}, the second step is triangle inequality, the third and fourth steps are due to the submultiplicative norm property, the fifth step comes from the definition of the Frobenius norm, the fact that M_{pre} and Z_{pre} are both of dimension $n \times T_0$, and bounds on data entries. The final step collects terms.

Finally, we need only to obtain a bound on $\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2]$.

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{z_{pre}} - Z_{pre}^\top \boldsymbol{f}||_2] &\leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T_0}(z_t - Z_t^\top \boldsymbol{f})^2}\right] \\
&\overset{(a)}{\leq} \sqrt{\sum_{t=1}^{T_0}\mathbb{E}[(z_t - Z_t^\top \boldsymbol{f})^2]} \\
&= \sqrt{\sum_{t=1}^{T_0}\mathbb{E}[z_t^2 - 2z_t Z_t^\top \boldsymbol{f} + (Z_t^\top \boldsymbol{f})^2]} \\
&\overset{(b)}{=} \sqrt{\sum_{t=1}^{T_0}(\sigma^2 + \mathbb{E}[(Z_t^\top \boldsymbol{f})^2])} \\
&= \sqrt{T_0\sigma^2 + \sum_{t=1}^{T_0}\mathbb{E}[\sum_{i=1}^{n}(z_i f_i)^2])} \\
&\overset{(c)}{=} \sqrt{T_0\sigma^2 + \sum_{t=1}^{T_0}\sum_{i=1}^{n}\mathbb{E}[z_i^2 f_i^2])} \\
&= \sqrt{T_0\sigma^2 + \sum_{t=1}^{T_0}\sum_{i=1}^{n}\sigma^2 f_i^2} \\
&= \sqrt{T_0\sigma^2 + \sum_{t=1}^{T_0}\sigma^2||\boldsymbol{f}||_2^2} \\
&\overset{(d)}{\leq} \sqrt{T_0\sigma^2 + \sum_{t=1}^{T_0}\sigma^2} \\
&= \sqrt{2T_0\sigma^2}
\end{aligned}$$

Inequality (a) is due to Jensen's inequality. The step in (b) is because $\mathbb{E}[z_t Z_t^\top \boldsymbol{f}] = \mathbb{E}[z_t]\mathbb{E}[Z_t^\top \boldsymbol{f}] = 0$ by independence of noise terms. The step in (c) is by the same logic as in (b), since all cross-terms $f_i f_j$ for $i \neq j$ are zero in expectation. Lastly, we bound the ℓ_2 norm of $\boldsymbol{f}$ by ℓ_1 norm instead in (d) (i.e., $||\boldsymbol{f}||_2 \leq ||\boldsymbol{f}||_1 \leq 1$).

Hence,

$$\begin{aligned}
\mathbb{E}[||X_{pre}\boldsymbol{z_{pre}} - X_{pre}Z_{pre}^\top \boldsymbol{f}||_2 + \lambda] &\leq (\sqrt{nT_0} + \sqrt{nT_0 s^2})\sqrt{2T_0\sigma^2} + \lambda \\
&= T_0\sqrt{2n\sigma^2} + T_0\sqrt{2n\sigma^2 s^2} + \lambda
\end{aligned}$$

Finally, combining this with Equation (27) gives the desired bound to complete the proof of Lemma C.6.

$$\mathbb{E}[||\boldsymbol{f}^{reg} - \boldsymbol{f}||_2] \leq \frac{(\sqrt{2n\sigma^2} + \sqrt{2n\sigma^2 s^2})T_0 + \lambda}{(1-\xi)T_0 + \lambda}.$$

G Omitted Proofs for $DPSC_{obj}$

G.1 Proof of Lemma D.3

Lemma D.3 *For any* $\Delta \geq 0$, $\Phi(\boldsymbol{\alpha}; \Delta) = \frac{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))|}{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))|} \leq (1 + \frac{c}{\lambda + \Delta})^2$.

Recall that $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ is the noise value that must have been realized when database $\mathcal{D}$ was input and $\boldsymbol{\alpha} = \arg\min_{\boldsymbol{f}} J^{obj}(\boldsymbol{f})$ was output. Since $J^{obj}(\boldsymbol{f})$ is strongly convex for any Δ and is differentiable, the closed-form expression for $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ is derived by computing the gradient of $J^{obj}(\boldsymbol{f})$, which should be zero when evaluated at its minimizer $\boldsymbol{f} = \boldsymbol{\alpha}$:

$$\nabla J^{obj}(\boldsymbol{f})\big|_{\boldsymbol{f}=\boldsymbol{\alpha}} = \nabla \mathcal{L}(\boldsymbol{\alpha}) + \nabla r(\boldsymbol{\alpha}) + \frac{\Delta}{T_0}\boldsymbol{\alpha} + \frac{\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})}{T_0} \stackrel{!}{=} 0.$$

Rearranging the equation yields

$$\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}) = -\left(T_0 \nabla \mathcal{L}(\boldsymbol{\alpha}; \mathcal{D}) + T_0 \nabla r(\boldsymbol{\alpha}) + \Delta \boldsymbol{\alpha}\right).$$

For ease of notation, let $A = -\nabla \boldsymbol{b}(\alpha; \mathcal{D})$ and $E = \nabla \boldsymbol{b}(\alpha; \mathcal{D}) - \nabla \boldsymbol{b}(\alpha; \mathcal{D}')$. Then,

$$\Phi(\boldsymbol{\alpha}; \Delta) = \frac{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))|}{|det(\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))|} = \frac{|det(-\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}'))|}{|det(-\nabla \boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}))|} = \frac{|det(A+E)|}{|det(A)|}.$$

By definition, $A = -\nabla \boldsymbol{b}(\alpha; \mathcal{D}) = T_0(\nabla^2 \mathcal{L}(\alpha; \mathcal{D}) + \nabla^2 r(\alpha)) + \Delta I_n$. Using the Hessians $\nabla^2 \mathcal{L}(\alpha; \mathcal{D}) = \frac{2}{T_0} X_{pre} X_{pre}^\top$ and $\nabla^2 r(\alpha) = \frac{\lambda}{T_0} I_n$, A can be expressed as

$$A = 2X_{pre}X_{pre}^T + (\lambda + \Delta) I_n.$$

To express E succinctly, let neighboring databases $\mathcal{D} = (X, y)$ and $\mathcal{D}' = (X', y)$ differ in the j-th row. Then,

$$E = 2(X'_{pre}X'^\top_{pre} - X_{pre}X_{pre}^\top) = \begin{cases} 2(||\boldsymbol{x'_j}||_2^2 - ||\boldsymbol{x_j}||_2^2) & (j,j) \\ 2(\boldsymbol{x'_j} - \boldsymbol{x_j})^\top \boldsymbol{x_i} & (j,i) \text{ or } (i,j)\ ,\ \forall i \in [n],\ i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

where $\boldsymbol{x_i}$ (resp. $\boldsymbol{x'_i}$) denotes the i-th person's data, which is the i-th row of X_{pre} (resp. X'_{pre}).

Note that all eigenvalues of A are at least $\lambda + \Delta > 0$ (i.e., $\lambda_{min}(A) \geq \lambda + \Delta$) because $X_{pre}X_{pre}^\top$ is positive-semi-definite, and thus A is full rank. Also, $rank(E) = 2$. This allows us to apply the following lemma.

Lemma G.1 (Lemma 2 of Chaudhuri et al. (2011)) *If A is full rank and E has rank at most 2,*

$$\frac{det(A+E) - det(A)}{det(A)} = \lambda_1(A^{-1}E) + \lambda_2(A^{-1}E) + \lambda_1(A^{-1}E)\lambda_2(A^{-1}E),$$

where $\lambda_i(Z)$ is i-th eigenvalue of matrix Z.

Let $\lambda_{|max|}(Z) = \max_i |\lambda_i(Z)|$, the maximum absolute of eigenvalue of matrix Z. Instantiating Lemma G.1 yields:

$$\begin{aligned}
\Phi(\boldsymbol{\alpha}; \Delta) &= \frac{|det(A+E)|}{|det(A)|} \\
&= \left| \frac{det(A+E) - det(A)}{det(A)} + 1 \right| \\
&= |1 + \lambda_1(A^{-1}E) + \lambda_2(A^{-1}E) + \lambda_1(A^{-1}E)\lambda_2(A^{-1}E)| \\
&\leq 1 + |\lambda_1(A^{-1}E)| + |\lambda_2(A^{-1}E)| + |\lambda_1(A^{-1}E)\lambda_2(A^{-1}E)| \\
&\leq 1 + 2\lambda_{|max|}(A^{-1}E) + \lambda_{|max|}(A^{-1}E)^2,
\end{aligned}$$

where the first inequality is simply triangle inequality, and the second inequality bounds all absolute eigenvalues by the maximum one $\lambda_{|max|}$.

Assume that $\lambda_{|max|}(E) \leq c$ for some constant c. Since E is a real-valued matrix, such a finite c exist. In Algorithm 3, c is explicitly taken as an input parameter. Then,

$$\lambda_{|max|}(A^{-1}E) \leq \frac{\lambda_{|max|}(E)}{\lambda_{min}(A)} \leq \frac{c}{\lambda + \Delta}.$$

Finally,

$$\Phi(\boldsymbol{\alpha}; \Delta) \leq 1 + 2\lambda_{|max|}(A^{-1}E) + \lambda_{|max|}(A^{-1}E)^2 \leq 1 + \frac{2c}{\lambda+\Delta} + \frac{c^2}{(\lambda+\Delta)^2} \leq \left(1 + \frac{c}{\lambda+\Delta}\right)^2.$$

G.2 Proof of Lemma D.4

Lemma D.4 *When $\boldsymbol{b}$ is sampled according to pdf $p(\boldsymbol{b}; \beta) \propto \exp\left(-\frac{||\boldsymbol{b}||_2}{\beta}\right)$, where $\beta = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$, then $\Gamma(\boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}'))} \leq e^{\epsilon_0}$.*

We can start by re-writing $\Gamma(\boldsymbol{\alpha})$ as follows, where the first line directly comes from the pdf $\Pr(\boldsymbol{b}; \beta)$, the second line is due to reverse triangle inequality, and the third line is from the definition of $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ and canceling terms that occur in both $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D})$ and $\boldsymbol{b}(\boldsymbol{\alpha}; \mathcal{D}')$:

$$\begin{aligned}
\Gamma(\boldsymbol{\alpha}) &= \exp\left(-\frac{1}{\beta}\big|\, ||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D})||_2 - ||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}')||_2 \,\big|\right) \\
&\leq \exp\left(-\frac{1}{\beta}||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}) - \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}')||_2\right) \\
&= \exp\left(\frac{1}{\beta}||T_0\nabla\mathcal{L}(\boldsymbol{\alpha};\mathcal{D}') - T_0\nabla\mathcal{L}(\boldsymbol{\alpha};\mathcal{D})||_2\right). \qquad (29)
\end{aligned}$$

Next, we can continue to bound Equation (29) in two different ways, corresponding to the two possible values of β. The two values come from two different upper bounds on the sensitivity, and the minimum value will give a tighter bound.

The first upper bound uses Lemma B.4, and its notation of $g(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}, \mathcal{D}') - \mathcal{L}(\boldsymbol{f}, \mathcal{D})$ for neighboring databases $\mathcal{D}, \mathcal{D}'$. Then we can bound:

$$\begin{aligned}
(29) &\leq \exp\left(\frac{1}{\beta}||T_0\nabla g(\boldsymbol{\alpha})||_2\right) \\
&\leq \exp\left(\frac{1}{\beta}4T_0\sqrt{8+n}\right).
\end{aligned}$$

Hence, setting $\beta \geq \frac{4T_0\sqrt{8+n}}{\epsilon_0}$ makes $\Gamma(\alpha) \leq e^{\epsilon_0}$.

The second upper bound is based on c, and will yield a tighter bound when c is small. Recall that matrix E is defined in Equation (28), and that c is the upper bound $\boldsymbol{\lambda}_{|max|}(E) \leq c$. By plugging in $\nabla\mathcal{L}(\alpha) = \frac{1}{T_0}\left(2X_{pre}X_{pre}^\top\boldsymbol{\alpha} - 2X_{pre}\boldsymbol{y}_{pre}\right)$, we can alternatively bound:

$$\begin{aligned}
(29) &= \exp\left(\frac{1}{\beta}||2(X'_{pre}X'^\top_{pre} - X_{pre}X_{pre}^\top)\boldsymbol{\alpha} + 2(X_{pre} - X'_{pre})\boldsymbol{y}_{pre}||_2\right) \\
&\leq \exp\left(\frac{1}{\beta}||2(X'_{pre}X'^\top_{pre} - X_{pre}X_{pre}^\top)\boldsymbol{\alpha}||_2 + \frac{1}{\beta}||2(X'_{pre} - X_{pre})\boldsymbol{y}_{pre}||_2\right) \\
&\leq \exp\left(\frac{1}{\beta}||E\boldsymbol{\alpha}||_2 + \frac{4T_0}{\beta}\right) \\
&\leq \exp\left(\frac{1}{\beta}||E||_{2,2}||\boldsymbol{\alpha}||_2 + \frac{4T_0}{\beta}\right) \\
&\leq \exp\left(\frac{c\sqrt{n} + 4T_0}{\beta}\right),
\end{aligned}$$

where the second step is due to triangle inequality, the third step is plugging in the definition of E and bounding the second term based on the worst-case $X'_{pre} - X_{pre}$, which is all zeros with just one row with all 2's, and worst-case y_{pre}, which is all 1's). The fourth step is the submultiplicative property of operator norms, and the final step is due to the fact that $||E||_{2,2} = \boldsymbol{\lambda}_{|max|}(E) \leq c$ and that all elements of $\boldsymbol{\alpha} \in [-1, 1]^n$ are bounded by 1. Then setting $\beta \geq \frac{c\sqrt{n}+4T_0}{\epsilon_0}$ ensures $\Gamma(\alpha) \leq e^{\epsilon_0}$.

If either of the above conditions on β holds, then $\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_0}$ as desired. Thus we can choose $\beta = \min\{\frac{4T_0\sqrt{8+n}}{\epsilon_0}, \frac{c\sqrt{n}+4T_0}{\epsilon_0}\}$ that at least one will be satisfied. Taking the minimum rather than just one allows for a lower β and hence lower noise magnitude, while still satisfying the privacy requirement.

G.3 Proof of Lemma D.5

Lemma D.5 *When $\boldsymbol{b} \sim \mathcal{N}(0, \beta^2 I_n)$, where $\beta = \frac{4T_0\sqrt{8+n}\sqrt{2\log\frac{2}{\delta}+\epsilon_0}}{\epsilon_0}$, then $\Gamma(\boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}))}{\Pr(\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}'))} \leq e^{\epsilon_0}$ with probability at least $1 - \delta$.*

The proof of Lemma D.5 follows a similar structure to Lemma 14 of Kifer et al. (2012). We include the full proof for completeness. Let the noise term $\boldsymbol{b}$ be sampled from a multivariate Gaussian distribution $\mathcal{N}(0, \beta^2 I_n)$, and let $\mathcal{D}$ and $\mathcal{D}'$ be two arbitrary neighboring databases. Let $h(\boldsymbol{\alpha}) = \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}') - \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D})$ Then, we can express $\Gamma(\boldsymbol{\alpha})$ as,

$$\begin{aligned}
\Gamma(\boldsymbol{\alpha}) &= \frac{\exp(-\frac{||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D})||_2^2}{2\beta^2})}{\exp(-\frac{||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}')||_2^2}{2\beta^2})} \\
&= \exp(\frac{1}{2\beta^2}(||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}')||_2^2 - ||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D})||_2^2)) \\
&= \exp(\frac{1}{2\beta^2}(||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}) + h(\boldsymbol{\alpha})||_2^2 - ||\boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D})||_2^2)) \\
&= \exp(\frac{1}{2\beta^2}(2\langle \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}), h(\boldsymbol{\alpha})\rangle + ||h(\boldsymbol{\alpha})||_2^2)),
\end{aligned} \tag{30}$$

where the first step is from the distribution of noise $\boldsymbol{b}$, the final step is a binomial expansion applied to norms.

Note that,

$$\begin{aligned}
h(\boldsymbol{\alpha}) &= \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}') - \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}) \\
&= T_0(\nabla\mathcal{L}(\boldsymbol{\alpha};\mathcal{D}) - \nabla\mathcal{L}(\boldsymbol{\alpha};\mathcal{D}')) \\
&= -T_0\nabla g(\boldsymbol{\alpha}),
\end{aligned}$$

where $g(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha},\mathcal{D}') - \mathcal{L}(\boldsymbol{\alpha},\mathcal{D})$, as defined in Equation (5). By Lemma B.4, we know that $||\nabla g(\boldsymbol{\alpha})||_2 \leq 4\sqrt{8+n}$, so also

$$||h(\boldsymbol{\alpha})||_2 \leq 4T_0\sqrt{8+n}. \tag{31}$$

Similarly, because $\boldsymbol{b}$ is sampled from a multivariate Gaussian distribution $\mathcal{N}(0, \beta I_n)$ and sum of Gaussian variables is also Gaussian, then,

$$\langle \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}), h(\boldsymbol{\alpha})\rangle \ \sim \ \mathcal{N}(0, \beta^2||h(\boldsymbol{\alpha})||_2^2).$$

Since the exact distribution is known, we use a Gaussian tail bound to find a *well-behaving* set of $\boldsymbol{b}$.

Lemma G.2 (Chernoff bound for Gaussian Wainwright (2019)) *Let $Z \sim \mathcal{N}(0, \sigma^2)$. Then, for all $t > \sigma$,*

$$P[Z \geq t] \leq \exp(-\frac{t^2}{2\sigma^2}).$$

We instantiate Lemma G.2 with $Z = \langle \boldsymbol{b}(\alpha;\mathcal{D}), h(\boldsymbol{\alpha})\rangle$ and $t = \beta||h(\boldsymbol{\alpha})||_2\sqrt{2\log\frac{2}{\delta}}$. Note that $t > \sigma$ for any $\delta > 1/2$. Then,

$$\Pr\left[\langle \boldsymbol{b}(\alpha;\mathcal{D}), h(\boldsymbol{\alpha})\rangle \geq \beta||h(\boldsymbol{\alpha})||_2\sqrt{2\log\frac{2}{\delta}}\right] \leq \frac{\delta}{2},$$

which, by Equation (31) implies that,

$$\Pr\left[\langle \boldsymbol{b}(\alpha;\mathcal{D}), h(\boldsymbol{\alpha})\rangle \geq \beta(4T_0\sqrt{8+n})\sqrt{2\log\frac{2}{\delta}}\right] \leq \frac{\delta}{2}. \tag{32}$$

Define a set of values of $\boldsymbol{b}$, corresponding the the good event described by Equation (32): **GOOD** $= \{\boldsymbol{b} \mid \langle \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}), h(\boldsymbol{\alpha})\rangle \leq \beta(4T_0\sqrt{8+n})\sqrt{2\log\frac{2}{\delta}}\}$. By definition, $\Pr[\boldsymbol{b} \in \textbf{GOOD}] \geq 1-\delta$. That is, with probability at least $1-\delta$, the noise vector $\boldsymbol{b}$ is in the well-behaving set **GOOD**.

When $\boldsymbol{b} \in$ **GOOD**, then we can complete the bound on $\Gamma(\boldsymbol{\alpha})$ from Equation (30), combining the bound on $||h(\boldsymbol{\alpha})||_2^2$ from Equation (31):

$$\Gamma(\boldsymbol{\alpha}) = \exp(\frac{1}{2\beta^2}[2\langle \boldsymbol{b}(\boldsymbol{\alpha};\mathcal{D}), h(\boldsymbol{\alpha})\rangle + ||h(\boldsymbol{\alpha})||_2^2]) \leq \exp\left(\frac{1}{2\beta^2}[2\beta(4T_0\sqrt{8+n})\sqrt{2\log\frac{2}{\delta}} + (4T_0\sqrt{8+n})^2]\right).$$

Finally, the goal is to bound $\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_0}$, in the case where $\boldsymbol{b} \in$ **GOOD**. Solving the expression above for β yields

$$\begin{aligned}\beta &\geq \frac{1}{2}\left(\frac{(4T_0\sqrt{8+n})\sqrt{2\log\frac{2}{\delta}}}{\epsilon_0} + \sqrt{\frac{(4T_0\sqrt{8+n})^2 2\log\frac{2}{\delta}}{\epsilon_0^2} + \frac{(4T_0\sqrt{8+n})^2}{\epsilon_0}}\right) \\ &= \frac{1}{2}\left(\frac{4T_0\sqrt{8+n}}{\epsilon_0}\left(\sqrt{2\log\frac{2}{\delta}} + \sqrt{2\log\frac{2}{\delta} + \epsilon_0}\right)\right)\end{aligned} \tag{33}$$

Note that choosing

$$\beta \geq \frac{(4T_0\sqrt{8+n})\sqrt{2\log\frac{2}{\delta} + \epsilon_0}}{\epsilon_0}$$

satisfies the bound of Equation (33).

Thus $\Gamma(\boldsymbol{\alpha}) \leq e^{\epsilon_0}$, conditioned on $\boldsymbol{b} \in$ **GOOD**, which occurs with probability at least $1-\delta$.

G.4 Proof of Lemma E.1

Lemma E.1 *The ℓ_2 distance between $\boldsymbol{f}^{obj}$ and $\boldsymbol{f}^{reg}$ satisfies:*

$$\mathbb{E}[||\boldsymbol{f}^{obj} - \boldsymbol{f}^{reg}||_2] \leq \frac{2}{\lambda+\Delta}\mathbb{E}[||\boldsymbol{b}||_2] + \mathbb{1}_{\Delta\neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda+\Delta}\right)2T_0^2\sqrt{n},$$

where $\boldsymbol{b}$ and Δ are computed internally by Algorithm 3.

Recall the objective functions J^{obj} and J^{reg}:

$$J^{obj}(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}) + \frac{\lambda+\Delta}{2T_0}\|\boldsymbol{f}\|_2^2 + \frac{1}{T_0}\boldsymbol{b}^\top\boldsymbol{f} \quad \text{and} \quad J^{reg}(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}) + \frac{\lambda}{2T_0}\|\boldsymbol{f}\|_2^2,$$

with their respective minimizers $\boldsymbol{f}^{obj}$ and $\boldsymbol{f}^{reg}$. Define another objective function $J^{\#}$ and its minimizer $\boldsymbol{f}^{\#}$,

$$J^{\#}(\boldsymbol{f}) = \mathcal{L}(\boldsymbol{f}) + \frac{\lambda+\Delta}{2T_0}\|\boldsymbol{f}\|_2^2$$

which is a noise-free variant of J^{obj}.

We will express the difference between $\boldsymbol{f}^{reg}$ and $\boldsymbol{f}^{obj}$ using $\boldsymbol{f}^{\#}$ as an intermediate value:

$$\|\boldsymbol{f}^{reg} - \boldsymbol{f}^{obj}\|_2 = \|\boldsymbol{f}^{reg} - \boldsymbol{f}^{\#} + \boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2 \leq \|\boldsymbol{f}^{reg} - \boldsymbol{f}^{\#}\|_2 + \|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2. \tag{34}$$

We will bound these two terms separately, starting with $\|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2$. It is known that J^{obj} is $(\frac{\lambda+\Delta}{T_0})$-strongly convex, and that the gradient of of J^{obj} evaluated at its minimizer $\boldsymbol{f}^{obj}$ is zero. Then by the definition of strong convexity,

$$\|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2^2 \leq \left(J^{obj}(\boldsymbol{f}^{\#}) - J^{obj}(\boldsymbol{f}^{obj})\right) \frac{2T_0}{\lambda+\Delta}. \tag{35}$$

We can proceed to bound the difference in the objective function J^{obj} at these two points:

$$\begin{aligned} J^{obj}(\boldsymbol{f}^{\#}) - J^{obj}(\boldsymbol{f}^{obj}) &= \left(J^{\#}(\boldsymbol{f}^{\#}) + \frac{1}{T_0} b^\top \boldsymbol{f}^{\#}\right) - \left(J^{\#}(\boldsymbol{f}^{obj}) + \frac{1}{T_0} b^\top \boldsymbol{f}^{obj}\right) \\ &= \left(J^{\#}(\boldsymbol{f}^{\#}) - J^{\#}(\boldsymbol{f}^{obj})\right) + \left(\frac{1}{T_0} b^\top \boldsymbol{f}^{\#} - \frac{1}{T_0} b^\top \boldsymbol{f}^{obj}\right) \\ &\leq 0 + \frac{1}{T_0}\|\boldsymbol{b}\|_2 \|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2 \end{aligned}$$

where the inequality is due to the fact that $J^{\#}(\boldsymbol{f}^{\#}) \leq J^{\#}(\boldsymbol{f}^{obj})$, since $\boldsymbol{f}^{\#}$ is the minimizer of $J^{\#}$.

Plugging this into Equation (35) gives

$$\|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2^2 \leq \frac{1}{T_0}\|\boldsymbol{b}\|_2 \|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2 \frac{2T_0}{\lambda+\Delta},$$

or equivalently,

$$\|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2 \leq \|\boldsymbol{b}\|_2 \frac{2}{\lambda+\Delta}.$$

To bound the first term of Equation (34), we observe that if $\Delta = 0$, then $J^{\#} = J^{reg}$ and thus $\boldsymbol{f}^{\#} = \boldsymbol{f}^{reg}$, so $\|\boldsymbol{f}^{reg} - \boldsymbol{f}^{\#}\|_2 = 0$. Thus we only need to bound the distance when $\Delta \neq 0$.

We can write $\boldsymbol{f}^{reg}$ and $\boldsymbol{f}^{\#}$ using their closed-form expressions,

$$\boldsymbol{f}^{reg} = (X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0} I)^{-1} X_{pre}\boldsymbol{y_{pre}} \quad \text{and} \quad \boldsymbol{f}^{\#} = (X_{pre}X_{pre}^\top + \frac{\lambda+\Delta}{2T_0} I)^{-1} X_{pre}\boldsymbol{y_{pre}},$$

and use these to bound the difference:

$$\begin{aligned} \|\boldsymbol{f}^{reg} - \boldsymbol{f}^{\#}\|_2 &= \| \left((X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0} I)^{-1} - (X_{pre}X_{pre}^\top + \frac{\lambda+\Delta}{2T_0} I)^{-1}\right) X_{pre}\boldsymbol{y_{pre}}\|_2 \\ &\leq \left((\|(X_{pre}X_{pre}^\top + \frac{\lambda}{2T_0} I)^{-1}\|_2 + \|((X_{pre}X_{pre}^\top + \frac{\lambda+\Delta}{2T_0} I)^{-1})^{-1}\|_2 \right) \|X_{pre}\boldsymbol{y_{pre}}\|_2 \end{aligned} \tag{36}$$

The spectral norm of a general form $\|(XX^\top + \lambda I)^{-1}\|_2$ can be bounded by the inverse of minimum singular value of the matrix $XX^\top + \lambda I$, which is positive semi-definite and has minimum singular value at least λ:

$$\|(XX^\top + \lambda I)^{-1}\|_2 \leq \frac{1}{\boldsymbol{\sigma_{min}}(XX^\top + \lambda I)} \leq \frac{1}{\lambda}.$$

Using this fact, we can further bound Equation (36) as,

$$\begin{aligned} \|\boldsymbol{f}^{reg} - \boldsymbol{f}^{\#}\|_2 &\leq \left(\frac{2T_0}{\lambda} + \frac{2T_0}{\lambda+\Delta}\right) \|X_{pre}\boldsymbol{y_{pre}}\|_2 \\ &\leq 2T_0 \left(\frac{1}{\lambda} + \frac{1}{\lambda+\Delta}\right) ||X_{pre}||_F ||\boldsymbol{y_{pre}}||_2 \\ &\leq 2T_0 \left(\frac{1}{\lambda} + \frac{1}{\lambda+\Delta}\right) \sqrt{nT_0}\sqrt{T_0} \\ &= \left(\frac{1}{\lambda} + \frac{1}{\lambda+\Delta}\right) 2T_0^2\sqrt{n}. \end{aligned}$$

Finally, we combine Equation (34) with bounds on both terms to yield:

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{f}^{reg} - \boldsymbol{f}^{obj}\|_2] &\leq \mathbb{E}[\|\boldsymbol{f}^{\#} - \boldsymbol{f}^{obj}\|_2] + \mathbb{1}_{\Delta\neq 0}\mathbb{E}[\|\boldsymbol{f}^{reg} - \boldsymbol{f}^{\#}||_2] \\ &\leq \frac{2}{\lambda+\Delta}\mathbb{E}[\|\boldsymbol{b}\|_2] + \mathbb{1}_{\Delta\neq 0}\left(\frac{1}{\lambda} + \frac{1}{\lambda+\Delta}\right) 2T_0^2\sqrt{n}. \end{aligned}$$