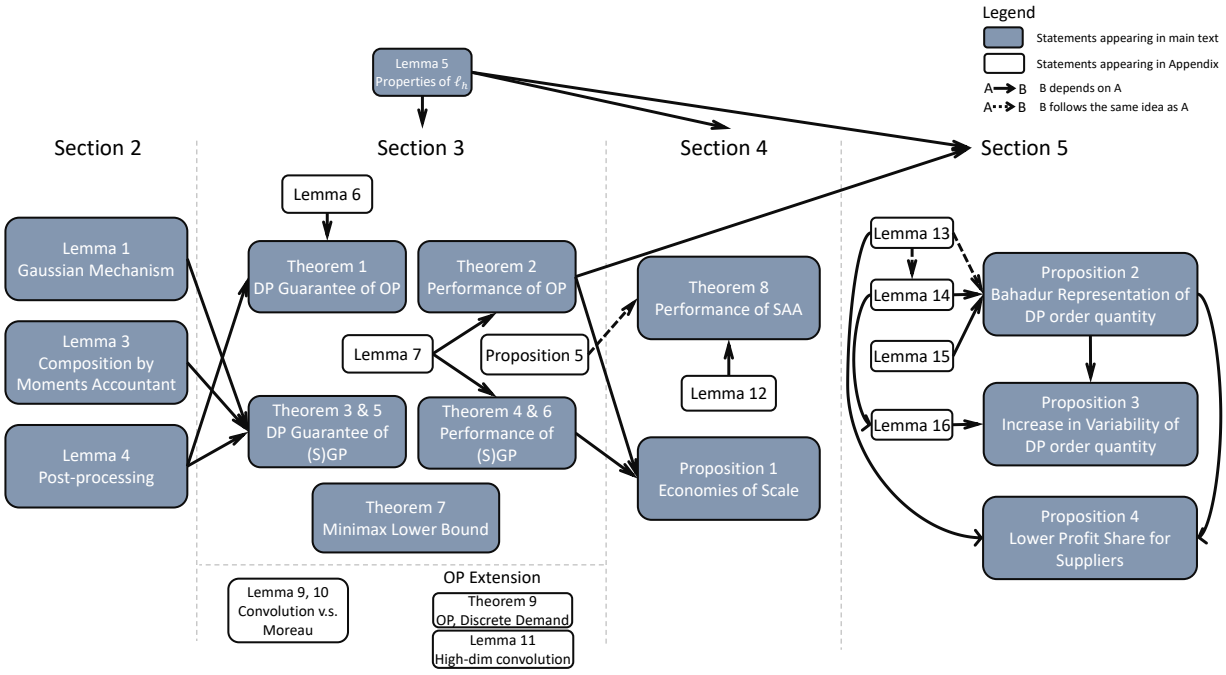
This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

E-Companion

We would like to highlight a remark regarding notations before presenting the proofs. For the proofs in Appendix, we will frequently adopt $\widehat{\mathcal{L}}(\boldsymbol{\theta})$ as a shorthand for $\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D})$ *when no manipulations on dataset* $\mathcal{D}$. Always keep in mind that the hat symbol $\widehat{\ }$ reflects implicitly the dependence on the underlying dataset $\mathcal{D}$. We also adopt the same notation for the derivatives $\nabla\widehat{\mathcal{L}}(\boldsymbol{\theta})$ and minimizers $\widehat{\boldsymbol{\theta}}$. But when analysis involves manipulating dataset $\mathcal{D}$, we will explicitly express such dependence. Additionally, $\mathcal{L}(\boldsymbol{\theta})$ is the shorthand for population-level function $\mathcal{L}(\boldsymbol{\theta};\mathbb{P})$, and the same rules are applied to population-level derivatives $\nabla\mathcal{L}(\boldsymbol{\theta})$ and minimizers $\boldsymbol{\theta}^*$.

Figure EC.1 below is the dependence diagram of proofs.

Figure EC.1 dependence diagram of proofs.



Note. Boxes with(out) backgrounds are statements appearing in the main text (appendix). Solid (dashed) lines indicate strong (weak) dependence between proofs.

EC.1. Proofs for Section 3

EC.1.1. Proof of Lemma 5

Proof.

- Gradients in part 1 of this Lemma can be easily calculated from function $c_h(\cdot)$.
- As for part 2, L-Lipschitz continuity is by noticing $\sup_{\boldsymbol{\theta},\boldsymbol{x},y}\|\nabla\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\|_2=\bar{r}B_x$; β-smooth is by noticing the supremum of maximal eigenvalue of Hessian matrix is $\sup_{\boldsymbol{\theta};\boldsymbol{x},y}\lambda_{\max}\left(\nabla^2\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\right)=\sup_{\boldsymbol{\theta};\boldsymbol{x},y}\lambda_{\max}\left(K_h\left(y-\boldsymbol{\theta}^\top\boldsymbol{x}\right)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right)\le\overline{K}B_x^2/h$.

- To prove part 3, we first note that because kernel function K is symmetric, the value $c_h(u)$ admits many equivalent formulations:

$$\begin{aligned} c_h(u) &= \frac{1}{2}\int_{-\infty}^{\infty} |u+hv|\,K(v)\,dv + (r-1/2)u && \text{(EC.1)}\\ &= \frac{1}{2}\int_{-\infty}^{\infty} |u-hv'|\,K(-v')\,dv' + (r-1/2)u && \text{(change variable } v' := -v\text{)}\\ &= \frac{1}{2}\int_{-\infty}^{\infty} |u-hv'|\,K(v')\,dv' + (r-1/2)u && \text{(kernel } K \text{ is symmetric)}\\ &= \frac{1}{2}\int_{-\infty}^{\infty} |u-hv|\,K(v)\,dv + (r-1/2)u. && \text{(EC.2)} \end{aligned}$$

Combining (EC.1) and (EC.2), we know

$$c_h(u) = \frac{1}{2}\int_{-\infty}^{\infty} \frac{|u+hv|+|u-hv|}{2} K(v)\,dv + (r-1/2)u.$$

To show $c_h(u) \geq c(u)$, it suffices to show $c_h(u) - c(u) \geq 0$. This is true, because

$$\begin{aligned} c_h(u) - c(u) &= \frac{1}{2}\int_{-\infty}^{\infty} \frac{|u+hv|+|u-hv|}{2} K(v)\,dv - \frac{|u|}{2}\\ &= \frac{1}{2}\int_{-\infty}^{\infty} \left(\frac{|u+hv|+|u-hv|}{2} - |u|\right) K(v)\,dv && \text{(EC.3)}\\ &\geq 0, \end{aligned}$$

where the last inequality is by convexity of the absolute function $|\cdot|$. Moreover, by symmetry of $|\cdot|$, the integrand in (EC.3) is strictly positive only in the set $\{v \in \mathbb{R} : |v| \geq |u|/h\}$. Consequently,

$$\begin{aligned} c_h(u) - c(u) &= \frac{1}{2}\int_{v:|v|\geq|u|/h} \left(\frac{|u+hv|+|u-hv|}{2} - |u|\right) K(v)\,dv\\ &\leq \frac{h}{2}\int_{v:|v|\geq|u|/h} |v|\,K(v)\,dv && \text{(triangular ineq.)}\\ &\leq \frac{h}{2}\int_{\mathbb{R}} |v|\,K(v)\,dv =: \frac{h}{2}\kappa_1, \end{aligned}$$

where the last two lines give upper bounds.

□

EC.1.2. Lemma EC.1 and Proof

While objective perturbation has been known for more than a decade, there is a technical bug in the original proof that has only been discovered recently (Agarwal et al. 2023, Redberg et al. 2024). In short, the bug appears at the top of page 25.21 in Kifer et al. (2012): "Note that Γ is independent of the noise vector." The independence claim is not well justified, leading to a wrong dependence on d in their conclusion. We follow the recent advancement to fix this bug for the private newsvendor problem.

Lemma EC.1 (Objective Perturbation for the Newsvendor Problem). *Let $\mathcal{D}=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\}$ be a dataset. Assume that the convex loss function can be written as $\ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i):=c_h(y_i-\langle\boldsymbol{x}_i,\boldsymbol{\theta}\rangle)$ for a known function $c_h(u)$ defined in (3). Let $\widehat{\mathcal{L}}_h(\boldsymbol{\theta};\mathcal{D})=\frac{1}{n}\sum_{i=1}^n\ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)$ be a loss function at dataset $\mathcal{D}$. Let $L:=\bar{r}B_x$ be the upper bound on $\|\nabla\ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\|_2$ and let β be an upper bound on the maximal eigenvalue of $\nabla^2\ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)$. Then $\widehat{\boldsymbol{\theta}}:=\arg\min\widehat{\mathcal{L}}_h(\boldsymbol{\theta};\mathcal{D})+\lambda\|\boldsymbol{\theta}\|_2+\boldsymbol{b}^\top\boldsymbol{\theta}/n$ is (ε,δ)-DP if*

1. *the noise vector $\boldsymbol{b}$ is sampled from a multivariate Gaussian $\mathcal{N}(0,\sigma^2\boldsymbol{I}_{d\times d})$ with variance $\sigma^2\geq L^2\cdot(8\ln(1/\delta)+4\varepsilon)/\varepsilon^2$;*
2. *the regularization coefficient $\lambda\geq\beta/(n\varepsilon)$.*

Proof. The proof follows the same proof idea for Theorem 2 in Kifer et al. (2012), with the technical bug being fixed. We will point out the bug and fix it in due course.

Let $\mathcal{A}(\mathcal{D}):=\arg\min_{\boldsymbol{\theta}}\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta};\mathcal{D})=\widehat{\mathcal{L}}_h(\boldsymbol{\theta};\mathcal{D})+\lambda\|\boldsymbol{\theta}\|_2^2+\frac{\boldsymbol{b}^\top\boldsymbol{\theta}}{n}$, and we are going to show, for any $\boldsymbol{v}$ and $\mathcal{D}\sim\mathcal{D}'$,

$$\frac{\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D})=\boldsymbol{v}]}{\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D}')=\boldsymbol{v}]}\leq e^{\varepsilon},\quad\text{w.p. at least } 1-\delta.$$

By first-order condition, we know that for any given $\mathcal{D}$, if the algorithm's output is $\mathcal{A}(\mathcal{D})$, then the noise drawn must be $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})=-n\nabla\widehat{\mathcal{L}}_h(\mathcal{A}(\mathcal{D});\mathcal{D})-2n\lambda\mathcal{A}(\mathcal{D})$. Changing variables according to the function inverse theorem A35 in Billingsley (2017), we can represent the output $\mathcal{A}(\mathcal{D})$ as a function of $\boldsymbol{b}$ in a probabilistic way; that is $\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D})=\boldsymbol{v}]=pdf(\boldsymbol{b}(\boldsymbol{v};\mathcal{D}))\cdot|\det(\nabla\boldsymbol{b}(\boldsymbol{v};\mathcal{D}))|$ for any possible output $\boldsymbol{v}$. Here, on the right-hand-side, $pdf(\boldsymbol{b}(\cdot;\mathcal{D}))$ is the pdf of noise $\boldsymbol{b}$, $\nabla\boldsymbol{b}$ is a function of $\boldsymbol{v}$, and $\det(\cdot)$ is the determinant of a given matrix. Therefore, when the output is $\mathcal{A}(\mathcal{D})=\boldsymbol{v}$, we must have

$$\frac{\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D})=\boldsymbol{v}]}{\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D}')=\boldsymbol{v}]}=\frac{pdf(\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}))}{pdf(\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}'))}\cdot\frac{|\det(\nabla\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}))|}{|\det(\nabla\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}'))|},\quad\forall\boldsymbol{v}.\tag{EC.4}$$

Without loss of generality, we assume $\mathcal{D}'$ has one more entry $(\boldsymbol{x}_n,y_n)$ than $\mathcal{D}$. We need to upper bound the two ratios on the r.h.s. of (EC.4).

We first check the ratio between the two pdfs, $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})$ and $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}')$. There is a clear relationship between the two random variables:

$$\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}')=\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})+\underbrace{\nabla\ell_h(\mathcal{A}(\mathcal{D});\boldsymbol{x}_n,y_n)}_{=:\Delta(\mathcal{A}(\mathcal{D}))}.$$

Remember that, the noise $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})$, and therefore, $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}')\sim\mathcal{N}(\Delta(\mathcal{A}(\mathcal{D})),\sigma^2\boldsymbol{I})$. Their likelihood ratio thus becomes

$$\begin{aligned}\frac{pdf(\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}))}{pdf(\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}'))}&=\frac{\exp\left(-\frac{1}{2}\|\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})\|_2^2/\sigma^2\right)}{\exp\left(-\frac{1}{2}\|\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})-\Delta(\mathcal{A}(\mathcal{D}))\|_2^2/\sigma^2\right)}\\&=\exp\left(\left[-\langle\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}),\Delta(\mathcal{A}(\mathcal{D}))\rangle+\frac{1}{2}\|\Delta(\mathcal{A}(\mathcal{D}))\|_2^2\right]/\sigma^2\right).\end{aligned}\tag{EC.5}$$

Through the analysis, we know $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})$ and $\Delta(\mathcal{A}(\mathcal{D}))$ are not independent, whereas Kifer et al. (2012) claims they are independent, which is incorrect. Since ℓ_h can be written as a linear function c with bounded derivatives, the norm of $\Delta(\mathcal{A}(\mathcal{D}))$ is bounded:

$$\|\Delta(\mathcal{A}(\mathcal{D}))\|_2 = \|\boldsymbol{x}_n \cdot c_h'(y_n - \langle \mathcal{A}(\mathcal{D}), \boldsymbol{x}_n\rangle)\|_2 \le \bar{r} B_x = L. \tag{EC.6}$$

The inequality in (EC.6) is because $c_h'(\cdot)$ is bounded by $\bar{r}$. Moreover, by $\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$,

$$\begin{aligned}\langle \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}), \Delta(\mathcal{A}(\mathcal{D}))\rangle &= \langle \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}), \boldsymbol{x}_n \cdot c_h'(y_n - \langle \mathcal{A}(\mathcal{D}), \boldsymbol{x}_n\rangle)\rangle \\ &= \langle \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}), \boldsymbol{x}_n\rangle \cdot c_h'(y_n - \langle \mathcal{A}(\mathcal{D}); \boldsymbol{x}_n\rangle) \\ &\sim \mathcal{N}(0, \|\boldsymbol{x}_n\|_2^2 \sigma^2) \cdot c_h'(y_n - \langle \mathcal{A}(\mathcal{D}); \boldsymbol{x}_n\rangle). \end{aligned} \tag{EC.7}$$

Since $c_h'(\cdot) \le \bar{r}$, it is immediate to conclude that the Gaussian random variable in (EC.7) has a lower variance than $\mathcal{N}(0, \bar{r}^2 B_x^2 \sigma^2) =: \mathcal{N}(0, L^2\sigma^2)$. Let us denote the relationship that B's variance is higher than A by $A \overset{var}{\le} B$. Then, (EC.7) implies

$$\langle \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}), \Delta(\mathcal{A}(\mathcal{D}))\rangle \overset{var}{\le} \mathcal{N}(0, L^2\sigma^2). \tag{EC.8}$$

Note that, with linear models, the reasoning for (EC.7) (EC.8) bypasses dependence issues in Kifer et al. (2012). Plugging (EC.6) and (EC.8) back into (EC.5), we get

$$\ln\left(\frac{pdf(\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}))}{pdf(\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}'))}\right) \overset{var}{\le} \left[\mathcal{N}(0, L^2\sigma^2) + L^2/2\right]/\sigma^2.$$

It remains to find a sufficiently large σ so that $[\mathcal{N}(0, L^2\sigma^2) + L^2/2]/\sigma^2 \le \frac{\varepsilon}{2}$ with probability at least $1-\delta$. By Gaussian random variable's tail bound $\Pr\left[\mathcal{N}(0,1^2) \ge \sqrt{2\ln(1/\delta)}\right] \le \delta$, it suffices to have $\frac{\varepsilon\sigma^2}{2L\sigma} - \frac{L^2}{2L\sigma} \ge \sqrt{2\ln(1/\delta)}$. Solving for σ, we get $\sigma \ge \frac{L\cdot\left(\sqrt{2\ln(1/\delta)}+\sqrt{2\ln(1/\delta)+\varepsilon}\right)}{\varepsilon}$. In other words, any σ greater than the stated value is sufficient for (ε,δ)-DP guarantee. We thus choose a slightly larger value:

$$\sigma = \frac{L\cdot\left(2\sqrt{2\ln(1/\delta)+\varepsilon}\right)}{\varepsilon} = \frac{L\cdot\sqrt{8\ln(1/\delta)+4\varepsilon}}{\varepsilon}.$$

Therefore, setting $\sigma^2 = \frac{L^2(8\ln(1/\delta)+4\varepsilon)}{\varepsilon^2}$ can ensure

$$\frac{pdf(\boldsymbol{b}(\boldsymbol{v};\mathcal{D}))}{pdf(\boldsymbol{b}(\boldsymbol{v};\mathcal{D}'))} \le e^{\frac{\varepsilon}{2}}, \quad \text{with prob. at least } 1-\delta. \tag{EC.9}$$

We then come to control the ratio between two determinants. The matrix $\nabla\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D})$ and $\nabla\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}')$ have a similar relationship:

$$\nabla\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}) = \nabla\boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}') + \underbrace{\nabla^2\ell_h(\mathcal{A}(\mathcal{D});\boldsymbol{x}_n, y_n)}_{=:E}.$$

For a given d-by-d symmetric matrix X, let $\rho_1(X) \geq \rho_2(X) \geq \cdots \geq \rho_d(X)$ be the eigenvalues of X. Specially, let $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_d \geq 0$ be the eigenvalues of $n\nabla^2 \widehat{\mathcal{L}}_h(\mathcal{A}(\mathcal{D});\mathcal{D})$, and let $\rho_1' \geq 0$ be the maximal eigenvalue of E since E is at most rank 1. Then, the ratio between determinants can be evaluated as:

$$\begin{aligned}
\frac{|\mathsf{det}(\nabla \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}'))|}{|\mathsf{det}(\nabla \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}))|} &= \frac{\left|\mathsf{det}(-n\nabla^2 \widehat{\mathcal{L}}_h(\mathcal{A}(\mathcal{D});\mathcal{D}) - 2n\lambda \boldsymbol{I} - E)\right|}{\left|\mathsf{det}(-n\nabla^2 \widehat{\mathcal{L}}_h(\mathcal{A}(\mathcal{D});\mathcal{D}) - 2n\lambda \boldsymbol{I})\right|} \\
&= \frac{\left|\Pi_{i=1}^d \rho_i(n\nabla^2 \widehat{\mathcal{L}}_h(\mathcal{A}(\mathcal{D});\mathcal{D}) + 2n\lambda \boldsymbol{I} + E)\right|}{\left|\Pi_{i=1}^d \rho_i(n\nabla^2 \widehat{\mathcal{L}}_h(\mathcal{A}(\mathcal{D});\mathcal{D}) + 2n\lambda \boldsymbol{I})\right|} \\
&\leq \frac{(\rho_1 + 2n\lambda + \rho_1')\,\Pi_{i=2}^d\,(\rho_i + 2n\lambda)}{\Pi_{i=1}^d\,(\rho_i + 2n\lambda)} \\
&= 1 + \frac{\rho_1'}{2n\lambda}.
\end{aligned}$$

The inequality is due to Weyl's theorems on eigenvalues for Hermitian matrices, which can be applied here because all involved matrices are positive semi-definite. Since ℓ_h is β-smooth, the eigenvalue ρ_1' should satisfy $\rho_1' \leq \beta$. To require

$$1 + \frac{\rho_1'}{2n\lambda} \leq 1 + \frac{\beta}{2n\lambda} \leq e^{\frac{\varepsilon}{2}}, \tag{EC.10}$$

it suffices to have $\lambda \geq \frac{\beta}{n\varepsilon}$. Therefore, we have

$$\frac{|\mathsf{det}(\nabla \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}'))|}{|\mathsf{det}(\nabla \boldsymbol{b}(\mathcal{A}(\mathcal{D});\mathcal{D}))|} \leq e^{\frac{\varepsilon}{2}}. \tag{EC.11}$$

Plugging (EC.9) and (EC.11) into (EC.4), we finally obtain that

$$\frac{\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D}) = \boldsymbol{v}]}{\Pr_{\mathcal{A}}[\mathcal{A}(\mathcal{D}') = \boldsymbol{v}]} \leq e^{\varepsilon}, \quad \text{with prob. at least } 1-\delta,$$

if $\sigma^2 \geq \frac{L^2 \cdot (8\ln(1/\delta) + 4\varepsilon)}{\varepsilon^2}$ and $\lambda \geq \frac{\beta}{n\varepsilon}$.

□

EC.1.3. Proof of Theorem 1

Proof of Theorem 1 According to Lemma EC.1, the minimizer $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$ of Objective Perturbation satisfies (ε,δ)-DP, because (i) noise vector $\boldsymbol{b}$ is sampled from a multivariate Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I}_{d\times d})$ with variance $\sigma^2 \geq (8\ln(1/\delta) + 4\varepsilon)/\varepsilon^2$; and (ii) the regularization coefficient $\lambda \geq \beta/(n\varepsilon)$. By Post-Processing Lemma 4, the final order quantity $\left\langle \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}, \boldsymbol{x}_{n+1} \right\rangle$ is therefore (ε,δ)-DP.

□

EC.1.4. Proof of Theorem 2

Table EC.1 below summarizes notations we are going to use for this proof.

Before proving Theorem 2, we first introduce the concept of Uniform Stability, and a related Lemma, which will help with our analysis.

Table EC.1 Notations for Objective Perturbation

Type	Description	Abbr.	Functions	Minimizers
Empirical	ERM	$\widehat{\mathcal{L}}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)$	$\widehat{\boldsymbol{\theta}}$
	Regularized ERM	$\widehat{\mathcal{L}}^{\#}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)+\lambda\left\|\boldsymbol{\theta}\right\|_2^2$	$\widehat{\boldsymbol{\theta}}^{\#}$
	Private regularized ERM	$\widehat{\mathcal{L}}^{\mathsf{OP}}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)+\lambda\left\|\boldsymbol{\theta}\right\|_2^2+\frac{\boldsymbol{b}^\top\boldsymbol{\theta}}{n}$	$\widehat{\boldsymbol{\theta}}^{\mathsf{OP}}$
Smoothed	ERM	$\widehat{\mathcal{L}}_h(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)$	$\widehat{\boldsymbol{\theta}}_h$
	Regularized ERM	$\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)+\lambda\left\|\boldsymbol{\theta}\right\|_2^2$	$\widehat{\boldsymbol{\theta}}_h^{\#}$
	Private regularized ERM	$\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell_h(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)+\lambda\left\|\boldsymbol{\theta}\right\|_2^2+\frac{\boldsymbol{b}^\top\boldsymbol{\theta}}{n}$	$\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$
Stochastic	Risk minimization	$\mathcal{L}(\boldsymbol{\theta})$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\ell(\boldsymbol{\theta};\boldsymbol{x},y)\right]$	$\boldsymbol{\theta}^*$
	Smoothed RM	$\mathcal{L}_h(\boldsymbol{\theta})$	$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\right]$	$\boldsymbol{\theta}_h^*$

Definition EC.1 (τ-Uniform Stability). *A randomized algorithm $\mathcal{A}:\mathcal{X}^n\times\mathcal{Y}^n\to\mathbb{R}^d$ is said to be τ-uniform stable with respect to function $\ell:\mathbb{R}^d\times\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ if for any pair of neighboring datasets $\mathcal{D}\sim\mathcal{D}'$ that differ in one data point only, we have*

$$\sup_{\boldsymbol{x},y}\mathbb{E}_{\mathcal{A}}\left[\ell(\mathcal{A}(\mathcal{D});\boldsymbol{x},y)-\ell(\mathcal{A}(\mathcal{D}');\boldsymbol{x},y)\right]\le\tau.$$

Lemma EC.2 (Uniform Stability Lemma). *(Bousquet and Elisseeff 2002) Let $\mathcal{A}:\mathcal{X}^n\times\mathcal{Y}^n\to\mathbb{R}^d$ be a τ-uniform stable algorithm w.r.t. loss function $\ell:\mathbb{R}^d\times\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. Let $\mathbb{P}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$, and $\mathcal{D}\sim\mathbb{P}^n$ be samples i.i.d. drawn from $\mathbb{P}$. Then, we have*

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n,\mathcal{A}}\left[\mathcal{L}(\mathcal{A}(\mathcal{D}))-\widehat{\mathcal{L}}(\mathcal{A}(\mathcal{D}))\right]\le\tau.$$

Proof of Theorem 2. We analyze the algorithm's performance without projection, which gives a looser upper bound. In the following analysis, we notationally suppress the dependencies on $\mathcal{D}$ and $\boldsymbol{b}$ unless explicitly manipulating them. Recall that the regret can be decomposed as

$$\mathcal{R}(\mathsf{OP};\mathbb{P})=\mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})-\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right]+\mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})-\widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right]+\left[\mathcal{L}_h(\boldsymbol{\theta}^*)-\mathcal{L}(\boldsymbol{\theta}^*)\right]. \tag{EC.12}$$

Thus, we can separately analyze the above three terms.

1) **(Uniform Stability)** We follow the standard stability analysis for strongly convex loss functions. For any $\mathcal{D}$ and $\boldsymbol{b}$, suppose that there is a dataset $\mathcal{D}'$ that differs from $\mathcal{D}$ in only one data point $(\boldsymbol{x}',y')$, then

$$\begin{aligned}
\lambda\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D})-\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')\right\|_2^2 &\le \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}');\mathcal{D})-\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}),\mathcal{D})\\
&=\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}');\mathcal{D}')-\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D});\mathcal{D}')\\
&\quad+\frac{\ell_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}');\boldsymbol{x},y)-\ell_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D});\boldsymbol{x},y)}{n}\\
&\quad+\frac{\ell_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D});\boldsymbol{x}',y')-\ell_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}');\boldsymbol{x}',y')}{n}\\
&\le\frac{2L\cdot\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D})-\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')\right\|_2}{n}.
\end{aligned} \tag{EC.13}$$

The first inequality is due to strong convexity of $\widehat{\mathcal{L}}_h^{\mathsf{OP}}$. The last inequality is due to the fact that $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')$ is the minimizer to $\min_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta};\mathcal{D}')$, and the fact that ℓ_h is L-Lipschitz continuous. The inequality (EC.13) implies that two OP minimizers trained on neighboring datasets are close to each other, i.e., $\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}) - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')\right\|_2 \leq \frac{2L}{n\lambda},\ \forall \mathcal{D}\sim\mathcal{D}', \boldsymbol{b}$. Recall that function $\ell(\cdot)$ is L-Lipschitz continuous in $\boldsymbol{\theta}$, thus

$$\ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D})) - \ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')) \leq L\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}) - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')\right\|_2 \leq \frac{2L^2}{n\lambda},\ \forall \mathcal{D}\sim\mathcal{D}', \boldsymbol{b}. \tag{EC.14}$$

Applying uniform stability Lemma EC.2, we get

$$\mathbb{E}_{\boldsymbol{b},\mathcal{D}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] \leq \frac{2L^2}{n\lambda}. \tag{EC.15}$$

2) **(Shrinking ERM)** We now fix a dataset $\mathcal{D}$ and a noise vector $\boldsymbol{b}$, and come to bound $\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)$. By strong convexity of $\widehat{\mathcal{L}}_h^{\mathsf{OP}}$ and Cauchy's inequality, we have

$$\lambda\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2 \leq \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) = \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) + \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\#}}{n} - \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}}{n} \leq \frac{\|\boldsymbol{b}\|_2\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2}{n},$$

which gives $\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2 \leq \frac{\|\boldsymbol{b}\|_2}{n\lambda}$. By relationships among minimizers, we further notice that

$$\begin{aligned}
\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*) &= \left[\widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \lambda\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2\right] - \left[\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta}^*) - \lambda\|\boldsymbol{\theta}^*\|_2^2\right] \\
&\leq \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\#}) + \lambda\left(\|\boldsymbol{\theta}^*\|_2^2 - \left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2\right) \\
&\leq \left[\widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] - \left[\widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\#})\right] + \lambda\|\boldsymbol{\theta}^*\|_2^2 \\
&= \left[\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}}{n}\right] - \left[\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\#}}{n}\right] + \lambda\|\boldsymbol{\theta}^*\|_2^2 \\
&\leq \left[\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}}{n}\right] - \left[\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\#}}{n}\right] + \lambda\|\boldsymbol{\theta}^*\|_2^2 \\
&\leq \frac{\|\boldsymbol{b}\|_2\cdot\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2}{n} + \lambda\|\boldsymbol{\theta}^*\|_2^2 \\
&\leq \frac{\|\boldsymbol{b}\|_2^2}{n^2\lambda} + \lambda\|\boldsymbol{\theta}^*\|_2^2,
\end{aligned} \tag{EC.16}$$

where the first inequality is from $\widehat{\boldsymbol{\theta}}_h^{\#}$ is the minimizer to $\widehat{\mathcal{L}}_h^{\#}$; the second inequality is by dropping a negative term $-\lambda\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2$; the third inequality is by replacing $\widehat{\boldsymbol{\theta}}_h^{\#}$ in the second bracket with minimizer $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$. The reasoning here holds for any $\mathcal{D}$ and $\boldsymbol{b}$, and therefore taking expectation over $\mathcal{D}$ and $\boldsymbol{b}$ yields

$$\mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] \leq \frac{\mathbb{E}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_2^2\right]}{n^2\lambda} + \lambda\|\boldsymbol{\theta}^*\|_2^2 \leq \frac{d\sigma^2}{n^2\lambda} + \lambda B_\theta^2. \tag{EC.17}$$

3) **(Convolution approximation error)** We first notice that the population-level approximation error is

$$\begin{aligned}
\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[c_h\left(y - \boldsymbol{\theta}^{*\top}\boldsymbol{x}\right) - c\left(y - \boldsymbol{\theta}^{*\top}\boldsymbol{x}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon^\star(\boldsymbol{x})}\left[c_h(\boldsymbol{\theta}^{\star\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}) - \boldsymbol{\theta}^{*\top}\boldsymbol{x}) - c(\boldsymbol{\theta}^{\star\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}) - \boldsymbol{\theta}^{*\top}\boldsymbol{x})\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon^\star(\boldsymbol{x})}\left[c_h(\boldsymbol{\delta}^{\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x})) - c(\boldsymbol{\delta}^{\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}))\right]\right] && (\text{let } \boldsymbol{\delta} := \boldsymbol{\theta}^\star - \boldsymbol{\theta}^*) \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon^\star(\boldsymbol{x})}\left[g_h(\boldsymbol{\delta}^{\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}))\right]\right]. && (\text{by definition of } g_h(\cdot))
\end{aligned}$$

From the proof of part 3 of Lemma 5, we know that function $g_h(u)$ can be upper bounded as:

$$g_h(u) \le h\int_{|u|/h}^{\infty} vK(v)\,dv, \quad \forall u \in \mathbb{R}.$$

Therefore, the population-level convolution approximation error can be upper bounded as, $\forall \boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}_{\epsilon^\star(\boldsymbol{x})}\left[g_h(\boldsymbol{\delta}^{\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}))\right] &\le h\int_{-\infty}^{\infty}\int_{\left|\boldsymbol{\delta}^{\top}\boldsymbol{x}+u\right|/h}^{\infty} vK(v)\,dv\,dF_{\epsilon^\star|\boldsymbol{x}}(u) \\
&= h\int_{0}^{\infty}\int_{-\boldsymbol{\delta}^{\top}\boldsymbol{x}-vh}^{-\boldsymbol{\delta}^{\top}\boldsymbol{x}+vh} vK(v)\,dF_{\epsilon^\star|\boldsymbol{x}}(u)\,dv && (\text{by Fubini's Theorem}) \\
&= h\int_{0}^{\infty} vK(v)\cdot\left(F_{\epsilon^\star|\boldsymbol{x}}(-\boldsymbol{\delta}^{\top}\boldsymbol{x} + vh) - F_{\epsilon^\star|\boldsymbol{x}}(-\boldsymbol{\delta}^{\top}\boldsymbol{x} - vh)\right)dv.
\end{aligned}$$

Since we assume $\epsilon^\star(\boldsymbol{x})$'s CDF $F_{\epsilon^\star|\boldsymbol{x}}$ is s-Lipschitz continuous, we have $F_{\epsilon^\star|\boldsymbol{x}}(-\boldsymbol{\delta}^{\top}\boldsymbol{x} + vh) - F_{\epsilon^\star|\boldsymbol{x}}(-\boldsymbol{\delta}^{\top}\boldsymbol{x} - vh) \le svh$. And therefore,

$$\begin{aligned}
\mathbb{E}_{\epsilon^\star(\boldsymbol{x})}\left[g_h(\boldsymbol{\delta}^{\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}))\right] &\le h^2\cdot 2s\int_{0}^{\infty} v^2K(v)\,dv \\
&= h^2\cdot s\int_{-\infty}^{\infty}|v|^2\,K(v)\,dv =: h^2\cdot M, \quad \forall h > 0,
\end{aligned}$$

where $M := s\kappa_2$ is a constant that depends on the Lipschitz parameter of $F_{\epsilon^\star|\boldsymbol{x}}$ and kernel function $K(\cdot)$. Since the upper bound $h^2\cdot M$ holds for any $\boldsymbol{x} \in \mathcal{X}$, the overall convolution approximation error $\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*)$ is upper bounded as

$$\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon^\star(\boldsymbol{x})}\left[g_h(\boldsymbol{\delta}^{\top}\boldsymbol{x} + \epsilon^\star(\boldsymbol{x}))\right]\right] \le \mathbb{E}_{\boldsymbol{x}}\left[h^2M\right] = h^2M. \tag{EC.18}$$

Lastly, combining the three parts together and plugging (EC.15), (EC.17), and (EC.18) into (EC.12), we get

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \le \frac{2L^2}{n\lambda} + \frac{d\sigma^2}{n^2\lambda} + \lambda B_\theta^2 + h^2M. \tag{EC.19}$$

Set $\lambda = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}}$ and $h = \frac{\overline{K}B_x^2}{\lambda n\varepsilon}$, we get

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \le 2B_\theta\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}} + \frac{\overline{K}^2B_x^4MB_\theta^2}{2L^2n\varepsilon^2 + d\sigma^2\varepsilon^2}.$$

To ensure the first term on the r.h.s. dominates the second term, it suffices to have $2B_\theta\sqrt{\frac{2L^2}{n}+\frac{d\sigma^2}{n^2}} \geq \frac{\overline{K}^2 B_x^4 M B_\theta^2}{2L^2 n\varepsilon^2+d\sigma^2\varepsilon^2}$. A sufficient condition to make it true is $\varepsilon \geq \frac{\overline{K}s\kappa_2}{2\sqrt{\bar{r}^3}}\cdot\sqrt{\frac{B_\theta}{B_x}}\cdot\frac{1}{\sqrt[4]{n}}$. Consequently,

$$\begin{aligned}
\mathcal{R}(\mathsf{OP};\mathbb{P}) &\leq 4B_\theta\sqrt{\frac{2L^2}{n}+\frac{d\sigma^2}{n^2}} \\
&= 4B_\theta B_x\bar{r}\sqrt{\frac{2}{n}+\frac{d(8\ln(1/\delta)+4\varepsilon)}{n^2\varepsilon^2}} \\
&\leq 4B_\theta B_x\bar{r}\sqrt{\frac{2}{n}+\frac{16d\ln(1/\delta)}{n^2\varepsilon^2}} \qquad \text{(since } \delta\leq e^{-\varepsilon/2}\Leftrightarrow 4\varepsilon\leq 8\ln(1/\delta)\text{)}.
\end{aligned}$$

The desired Theorem then follows by noticing that the above analysis can be applied to any distribution $\mathbb{P}$ that satisfies stated assumptions.

□

EC.1.5. Proof of Theorem 3

Lemma EC.3 (Advanced Composition). *(Dwork et al. 2010) For all $\varepsilon,\delta,\delta'\geq 0$, the class of (ε,δ)-DP mechanisms satisfies $(\varepsilon\sqrt{2k\ln(1/\delta')}+k\varepsilon(e^\varepsilon-1), k\delta+\delta')$-DP under k-fold adaptive composition.*

Proof. We start by using Gaussian Mechanism (Lemma 1) and Advanced Composition Lemma (Lemma EC.3) to derive a variance that nearly has the same magnitude as stated. We notice that, for any $\boldsymbol{\theta}\in\mathbb{R}^d$, the Global Sensitivity of the empirical gradient function $\nabla\widehat{\mathcal{L}}_h$ is

$$\mathsf{GS} = \sup_{\mathcal{D}\sim\mathcal{D}'}\left\|\nabla\widehat{\mathcal{L}}_h(\boldsymbol{\theta},\mathcal{D})-\nabla\widehat{\mathcal{L}}_h(\boldsymbol{\theta},\mathcal{D}')\right\|_2 \leq \frac{2}{n}\sup_{\boldsymbol{x},y}\|\nabla\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\|_2 \leq \frac{2L}{n}. \tag{EC.20}$$

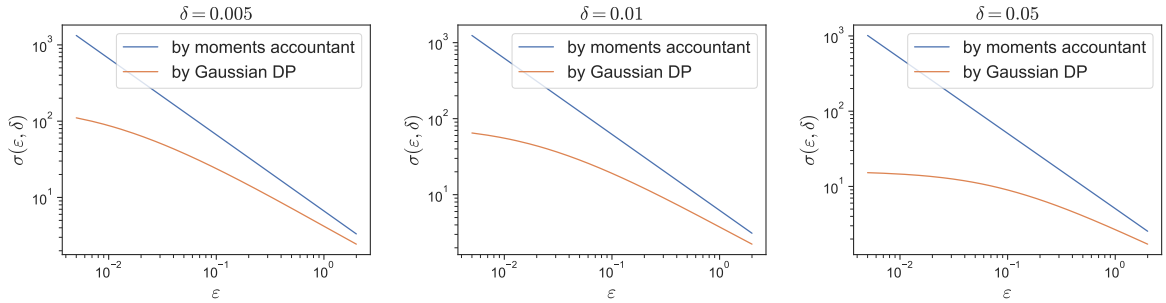
Therefore, by Gaussian Mechanism (Dwork and Roth 2014, Theorem A.1), when the variance of injected noise is $\sigma^2=\frac{2L^2\ln(1.25/\delta_0)}{n^2\varepsilon_0^2}$, it ensures (ε_0,δ_0)-DP for each iteration. By Advanced Composition Lemma (Lemma EC.3), to ensure an adaptive algorithm with T iterations to be (ε,δ)-DP, it suffices to set $\varepsilon_0=\frac{\varepsilon}{2\sqrt{2T\ln(2/\delta)}}$ and $\delta_0=\frac{\delta}{2T}$. Substituting these values into the expression of σ^2, we obtain $\sigma^2=\frac{64TL^2\ln(2.5T/\delta)\ln(2/\delta)}{n^2\varepsilon^2}$. The derived variance σ^2 nearly matches the magnitude of that of injected noise in Algorithm GP, but the constant and logarithmic terms show room for improvement.

To get rid of the dependence on $\ln T$ and further reduce the constant from 64 to 8, we can utilize Moments Accountant method (Abadi et al. 2016) or Rényi differential privacy (Mironov 2017) to track privacy loss more carefully, see for example Theorem 1 in Abadi et al. (2016), where they show that $\sigma^2=\frac{2\mathsf{GS}^2T\ln(1/\delta)}{\varepsilon^2}$ is sufficient for (ε,δ)-DP. Substituting GS with $\frac{2L}{n}$ gives the statement.

The recent advancements in privacy accounting techniques found that both Advanced Composition and Moments Accountant are informationally lossy for compositions of Gaussian Mechanisms, meaning that σ^2 suggested by them is unnecessarily large. Composition Lemma by Gaussian DP (Lemma 2) tightens the accounting and gives the smallest $\sigma(\varepsilon,\delta):=\inf\{\sigma>0:\delta(\sigma,\varepsilon)\leq\delta\}$ with $\delta(\sigma,\varepsilon)=\Phi\left(-\frac{\varepsilon\sigma}{\sqrt{T}\cdot\mathsf{GS}}+\frac{\sqrt{T}\cdot\mathsf{GS}}{2\sigma}\right)-$

$e^{\varepsilon}\Phi\left(-\frac{\varepsilon\sigma}{\sqrt{T}\cdot\mathsf{GS}}-\frac{\sqrt{T}\cdot\mathsf{GS}}{2\sigma}\right)$. To see the improvement, we draw σ versus ε under various σ in Figure EC.2. Evidently, the noise given by Gaussian-DP is much less than that by moments accountant. For details of Gaussian DP and broadly f-DP, interested readers are encouraged to refer to Dong et al. (2022). Nevertheless, because the value of $\sigma(\varepsilon,\delta)$ by Composition Lemma through Gaussian-DP does not admit a closed-form expression, we will stick to the closed-form variance given by moments accountant. Having a closed-form expression allows us to establish a clear understanding of the relationship between the price of privacy and the parameters (ε,δ).

Figure EC.2 Noise level suggested by different accounting techniques for (ε,δ)-DP. δ takes values as indicated in the subplots' title. T is set to be equal to n^2. For all curves, the lower the better.



□

EC.1.6. Proof of Theorem 4

Proof. We analyze the algorithm's performance without projection, which gives a looser upper bound. We follow *uniform stability* and *shrinking ERM* framework to complete our proof. The regret can be decomposed into two parts by injecting the term $\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}})$:

$$\mathcal{R}(\mathsf{GP};\mathbb{P}) = \mathbb{E}_{\mathsf{GP},\mathcal{D}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}})\right] + \mathbb{E}_{\mathsf{GP},\mathcal{D}}\left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right], \tag{EC.21}$$

where the expectation $\mathbb{E}_{\mathsf{GP}}[\cdot]$ means taking expectation over algorithm's randomness, i.e., sampling noise vectors $\{\boldsymbol{w}_t\}_{t=1}^T$.

When deriving the upper bound for the first term in (EC.21), for notation brevity, the superscribe $^{\mathsf{GP}}$ will be omitted, and we use $\widehat{\boldsymbol{\theta}}_{h,t}$ and $\widehat{\boldsymbol{\theta}}'_{h,t}$ to represent vectors in t-th iteration trained on neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, respectively. Their different data point is indexed by k, and let $\mathcal{D}^{-k}$, $\mathcal{D}'^{-k}$ denote datasets with $n-1$ data points by removing k-th data point. So $\mathcal{D}^{-k}$ and $\mathcal{D}'^{-k}$ are identical. We first conjecture that, if step size $\eta \le 2/\beta$, then

$$\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2 \le \frac{2L\eta t}{n}, \quad \forall t = 1,\ldots,T. \tag{EC.22}$$

This conjecture can be proved by induction. When $t=1$, by the setting of initial points, obviously it is true. Then suppose it is true for t-th iteration, it remains to check $(t+1)$-th iteration. By the update rule,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}_{h,t+1}-\widehat{\boldsymbol{\theta}}'_{h,t+1}\right\|_2 &= \left\|\left(\widehat{\boldsymbol{\theta}}_{h,t}-\eta(\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t};\mathcal{D})+\boldsymbol{w}_t)\right)-\left(\widehat{\boldsymbol{\theta}}'_{h,t}-\eta(\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}'_{h,t};\mathcal{D}')+\boldsymbol{w}_t)\right)\right\|_2\\ &\le \left\|\left(\widehat{\boldsymbol{\theta}}_{h,t}-\eta\frac{(n-1)}{n}\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t};\mathcal{D}^{-k})\right)-\left(\widehat{\boldsymbol{\theta}}'_{h,t}-\eta\frac{n-1}{n}\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}'_{h,t};\mathcal{D}'^{-k})\right)\right\|_2\\ &\quad+\frac{\eta}{n}\left\|\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t};\boldsymbol{x}_k,y_k)-\nabla\ell_h(\widehat{\boldsymbol{\theta}}'_{h,t};\boldsymbol{x}'_k,y'_k)\right\|_2.\end{aligned}$$

Then, the first term can be upper bounded by the non-expansiveness property of gradient update rule from Lemma 3.6 in Hardt et al. (2016):

$$\left\|\left(\widehat{\boldsymbol{\theta}}_{h,t}-\eta\frac{(n-1)}{n}\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t};\mathcal{D}^{-k})\right)-\left(\widehat{\boldsymbol{\theta}}'_{h,t}-\eta\frac{n-1}{n}\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}'_{h,t};\mathcal{D}'^{-k})\right)\right\|_2\le\left\|\widehat{\boldsymbol{\theta}}_{h,t}-\widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2;$$

and the second term can be upper bounded by Lipschiz continuity arguments, i.e., $\frac{\eta}{n}\left\|\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t};\boldsymbol{x}_k,y_k)-\nabla\ell_h(\widehat{\boldsymbol{\theta}}'_{h,t};\boldsymbol{x}'_k,y'_k)\right\|_2\le\frac{2L\eta}{n}$. Consequently,

$$\left\|\widehat{\boldsymbol{\theta}}_{h,t+1}-\widehat{\boldsymbol{\theta}}'_{h,t+1}\right\|_2\le\left\|\widehat{\boldsymbol{\theta}}_{h,t}-\widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2+\frac{2L\eta}{n}\le\frac{2L\eta(t+1)}{n}.$$

Therefore, our conjecture (EC.22) is correct. With this result, we are able to establish uniform stability:

$$\ell(\widehat{\boldsymbol{\theta}}_h;\boldsymbol{x},y)-\ell(\widehat{\boldsymbol{\theta}}'_h,\boldsymbol{x},y)\le L\cdot\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\boldsymbol{\theta}}_{h,t}-\widehat{\boldsymbol{\theta}}'_{h,t}\right)\right\|_2\le\frac{L}{T}\sum_{t=1}^{T}\frac{2L\eta t}{n}=\frac{L^2\eta(1+T)}{n},\quad\forall\boldsymbol{x},y.$$

Hence, by uniform stability Lemma EC.2, we have

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n,\mathsf{GP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}})-\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}})\right]\le\frac{L^2\eta(1+T)}{n}.\tag{EC.23}$$

It remains to control the second term in (EC.21). Let us fix a dataset $\mathcal{D}$. By the relationship in Lemma 5 part 3 and convexity of $\widehat{\mathcal{L}}_h(\cdot)$, we can show that

$$\begin{aligned}\mathbb{E}_{\mathsf{GP}}\left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{GP}})-\widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right]&\le\mathbb{E}_{\mathsf{GP}}\left[\widehat{\mathcal{L}}_h\left(\frac{1}{T}\cdot\sum_{t=1}^{T}\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}}\right)-\widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right]+\frac{1}{2}h\kappa_1\\ &\le\frac{1}{T}\cdot\mathbb{E}_{\mathsf{GP}}\left[\sum_{t=1}^{T}\left(\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}})-\widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right)\right]+\frac{1}{2}h\kappa_1\\ &\le\frac{1}{T}\cdot\mathbb{E}_{\mathsf{GP}}\left[\sum_{t=1}^{T}\left\langle\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}}-\boldsymbol{\theta}^*,\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}})\right\rangle\right]+\frac{1}{2}h\kappa_1.\\ &=\frac{1}{T}\cdot\mathbb{E}_{\mathsf{GP}}\left[\sum_{t=1}^{T}\left\langle\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}}-\boldsymbol{\theta}^*,\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}})+\boldsymbol{w}_t\right\rangle\right]+\frac{1}{2}h\kappa_1\\ &\le\frac{1}{T}\cdot\mathbb{E}_{\mathsf{GP}}\left[\frac{\|\boldsymbol{\theta}^*\|_2^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\left\|\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}})+\boldsymbol{w}_t\right\|_2^2\right]+\frac{1}{2}h\kappa_1,\\ &\le\frac{1}{T}\cdot\mathbb{E}_{\mathsf{GP}}\left[\frac{\|\boldsymbol{\theta}^*\|_2^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\left\|\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{GP}})\right\|_2^2\right]+\frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}_{\mathsf{GP}}\left[\|\boldsymbol{w}_t\|_2^2\right]+\frac{1}{2}h\kappa_1,\\ &\le\frac{\|\boldsymbol{\theta}^*\|_2^2}{2T\eta}+\frac{\eta L^2}{2}+\frac{\eta d\sigma^2}{2}+\frac{1}{2}h\kappa_1,\end{aligned}\tag{EC.24}$$

where the second and third lines are due to convexity of $\widehat{\mathcal{L}}_h$; the fourth line comes from the fact that $\boldsymbol{w}_t$ is drawn from a zero-mean Gaussian distribution and is independent of $\widehat{\boldsymbol{\theta}}^{\mathsf{GP}}_{h,t}$; the fifth line follows from a classic gradient descent analysis (Shalev-Shwartz and Ben-David 2014, Lemma 14.1) and from the gradient descent update rule in our algorithm; the sixth line comes again from the zero-mean $\boldsymbol{w}_t$ being independent; the last line is due to L-Lipschitz continuity of $\widehat{\mathcal{L}}_h$ and the fact that $\left\|\frac{\boldsymbol{w}_t}{\sigma}\right\|_2^2$ is a χ^2 random variable with d-degree.

Thus, plugging (EC.23) and (EC.24) into (EC.21), and then letting $h = \eta\overline{K}B_x^2$ (which ensures $\eta\beta = 1$, and the prerequisite for our conjecture (EC.22) is thus met), we obtain

$$\mathcal{R}(\mathsf{GP};\mathbb{P}) \le \frac{L^2(1+T)\eta}{n} + \frac{\|\boldsymbol{\theta}^*\|_2^2}{2T\eta} + \frac{\eta L^2}{2} + \frac{\eta d\sigma^2}{2} + \frac{\eta\overline{K}B_x^2\kappa_1}{2}. \tag{EC.25}$$

It is safe to replace $\|\boldsymbol{\theta}^*\|_2^2$ with its upper bound B_θ^2. Then minimizing the r.h.s over η, we know that we should use $\eta = \sqrt{\frac{B_\theta^2}{2T}/\left(\frac{L^2(1+T)}{n} + \frac{L^2}{2} + \frac{d\sigma^2}{2} + \frac{\kappa_1\overline{K}B_x^2}{2}\right)}$, and replace T with n^2. Plugging these values and stated σ into (EC.25), we obtain

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{GP};\mathbb{P}) &\le B_\theta\sqrt{\frac{2L^2}{n} + \frac{8dL^2\ln(1/\delta)}{n^2\varepsilon^2} + \frac{L^2+\kappa_1\overline{K}B_x^2}{n^2} + \frac{2L^2}{n^3}} \\
&\le B_\theta B_x\bar{r}\sqrt{\frac{4}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2} + \frac{1+\kappa_1\overline{K}/\bar{r}^2}{n^2}} && \text{(by } L = B_x\bar{r}\text{)} \\
&\le B_\theta B_x\bar{r}\sqrt{\frac{8}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} && \text{(since } n \ge \frac{1+\kappa_1\overline{K}/\bar{r}^2}{4}\text{)} \\
&= \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right),
\end{aligned}$$

which completes the proof (The sup is by noticing that the above analysis applies to any distribution $\mathbb{P}$).

□

EC.1.7. Proof of Theorem 5

Proof. Our stochastic gradient perturbation is a variant of Algorithm 1 in Bassily et al. (2014) by removing a multiplicative factor n in their update rule. Therefore, the variance of the injected noise in our algorithm is $1/n^2$-times theirs and can still guarantee (ε,δ)-DP.

□

EC.1.8. Proof of Theorem 6

We analyze the algorithm's performance without projection, which gives a looser upper bound.

Proof. The proof follows exactly the same idea as the proof of Theorem 4. Therefore, we omit some details here and only highlight the key steps. We first decompose the regret into two parts,

$$\mathcal{R}(\mathsf{SGP};\mathbb{P}) = \mathbb{E}_{\mathsf{SGP},\mathcal{D}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}})\right] + \mathbb{E}_{\mathsf{SGP},\mathcal{D}}\left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right], \tag{EC.26}$$

where the expectation $\mathbb{E}_{\mathsf{SGP}}[\cdot]$ means taking expectation over algorithm's randomness. Note that there are two types of randomness, one randomness comes from sampling record $(\boldsymbol{x}_{(t)}, y_{(t)})$ for evaluating gradients, and another randomness is from sampling noise vector $\boldsymbol{w}_t$.

First, we bound the first term in (EC.26) by using the uniform stability Lemma EC.2. Because of L-Lipschitz continuity of ℓ, we have

$$\mathbb{E}_{\mathsf{SGP}}\left[\ell(\widehat{\boldsymbol{\theta}}^{\mathsf{SGP}}(\mathcal{D}); \boldsymbol{x}, y) - \ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}(\mathcal{D}'); \boldsymbol{x}, y)\right] \le L \cdot \mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}(\mathcal{D}) - \widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}(\mathcal{D}')\right\|_2\right], \quad \forall \boldsymbol{x}, y, \mathcal{D} \sim \mathcal{D}'.$$

Therefore, it suffices to control the expected deviation between the returned vectors trained on two neighboring datasets. For notation brevity, the superscript $^{\mathsf{SGP}}$ and the dependence on $\mathcal{D}$ are omitted throughout this proof, and we use $\widehat{\boldsymbol{\theta}}_{h,t}$ and $\widehat{\boldsymbol{\theta}}'_{h,t}$ to represent vectors in t-th iteration trained on $\mathcal{D}$ and $\mathcal{D}'$, respectively. Firstly, we conjecture that, when step size $\eta \le 2/\beta$, we have

$$\mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2\right] \le \frac{2L\eta t}{n}, \quad \forall t = 1, \ldots, n^2. \tag{EC.27}$$

This can be proved by induction. When $t = 1$, by the setting of initial points $\widehat{\boldsymbol{\theta}}_{h,1} = \widehat{\boldsymbol{\theta}}'_{h,1} = \mathbf{0}$, obviously it is true. Then suppose that (EC.27) is true for t-th iteration, it remains to check $(t+1)$-th iteration. Let us fix a sequence of noise vector $\{\boldsymbol{w}_t\}_{t=1}^T$, then,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}_{h,t+1} - \widehat{\boldsymbol{\theta}}'_{h,t+1}\right\|_2 &= \left\|\left(\widehat{\boldsymbol{\theta}}_{h,t} - \eta(\nabla \ell_h(\widehat{\boldsymbol{\theta}}_{h,t}; \boldsymbol{x}_{(t)}, y_{(t)}) + \boldsymbol{w}_t)\right) - \left(\widehat{\boldsymbol{\theta}}'_{h,t} - \eta(\nabla \ell_h(\widehat{\boldsymbol{\theta}}'_{h,t}; \boldsymbol{x}'_{(t)}, y'_{(t)}) + \boldsymbol{w}_t)\right)\right\|_2 \\ &= \left\|\left(\widehat{\boldsymbol{\theta}}_{h,t} - \eta \cdot \nabla \ell_h(\widehat{\boldsymbol{\theta}}_{h,t}; \boldsymbol{x}_{(t)}, y_{(t)})\right) - \left(\widehat{\boldsymbol{\theta}}'_{h,t} - \eta \cdot \nabla \ell_h(\widehat{\boldsymbol{\theta}}'_{h,t}; \boldsymbol{x}'_{(t)}, y'_{(t)})\right)\right\|_2. \end{aligned} \tag{EC.28}$$

Note that because $(\boldsymbol{x}_{(t)}, y_{(t)})$ and $(\boldsymbol{x}'_{(t)}, y'_{(t)})$ are uniformly drawn from $\mathcal{D}$ and $\mathcal{D}'$, it implies that, with probability $1/n$, $(\boldsymbol{x}_{(t)}, y_{(t)}) \ne (\boldsymbol{x}'_{(t)}, y'_{(t)})$; and with probability $1 - 1/n$, $(\boldsymbol{x}_{(t)}, y_{(t)}) = (\boldsymbol{x}'_{(t)}, y'_{(t)})$. When drawn points are not the same, (EC.28) $\le \left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2 + 2\eta L$ by triangular inequality. When drawn points are the same, the gradient update rule is 1-expansive, i.e., (EC.28) $\le \left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2$ (Hardt et al. 2016, Lemma 3.6). Consequently, taking expectation over the randomness of the algorithm, we have

$$\begin{aligned}\mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t+1} - \widehat{\boldsymbol{\theta}}'_{h,t+1}\right\|_2\right] &\le \left(1 - \frac{1}{n}\right)\mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2\right] + \frac{1}{n}\mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2 + 2\eta L\right] \\ &\le \mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2\right] + \frac{2\eta L}{n} \\ &\le \frac{2\eta L(t+1)}{n}.\end{aligned}$$

Therefore, by induction, our conjecture (EC.27) is true. Recall that final output $\widehat{\boldsymbol{\theta}}_h = \frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{\theta}}_{h,t}$ is the averaged vector over all iterations; thus,

$$\mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h - \widehat{\boldsymbol{\theta}}'_h\right\|_2\right] \le \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathsf{SGP}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2\right] \le \frac{1}{T}\sum_{t=1}^T \frac{2\eta L t}{n} = \frac{(1+T)\eta L}{n}.$$

By uniform stability Lemma EC.2, the above inequality implies:

$$\mathbb{E}_{\mathsf{SGP},\mathcal{D}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}})\right] \le \frac{(1+T)\eta L^2}{n}. \tag{EC.29}$$

Then, we come to bound the second term in (EC.26). By convexity of $\widehat{\mathcal{L}}_h(\cdot)$, when we fix algorithm's sampling noise randomness SGP^{sn} and sampling record randomness SGP^{sr}, following the first three lines for (EC.24), we can show that

$$\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*) \le \frac{1}{T}\sum_{t=1}^{T}\left[\left\langle \widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}} - \boldsymbol{\theta}^*, \nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}})\right\rangle\right] + \frac{1}{2}h\kappa_1. \tag{EC.30}$$

Since record $(\boldsymbol{x}_{(t)}, y_{(t)})$ is uniformly sampled over $\mathcal{D}$ and independent across iterations, the expected gradient over sampling record (sr) randomness is unbiased, i.e., the relationship $\mathbb{E}_{\mathsf{SGP}^{sr}}\left[\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}}; \boldsymbol{x}_{(t)}, y_{(t)})\right] = \nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}}), \forall t$ is true. Thus, substituting the relationship into (EC.30) gives

$$\begin{aligned}
\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*) &\le \mathbb{E}_{\mathsf{SGP}^{sr}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle \widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}} - \boldsymbol{\theta}^*, \nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}}; \boldsymbol{x}_{(t)}, y_{(t)})\right\rangle\right] + \frac{1}{2}h\kappa_1 \\
&\le \frac{1}{T}\cdot\mathbb{E}_{\mathsf{SGP}^{sr}}\left[\frac{\|\boldsymbol{\theta}^*\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left\|\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}}; \boldsymbol{x}_{(t)}, y_{(t)}) + \boldsymbol{w}_t\right\|_2^2\right] \\
&\quad - \mathbb{E}_{\mathsf{SGP}^{sr}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle \widehat{\boldsymbol{\theta}}_{h,t}^{\mathsf{SGP}} - \boldsymbol{\theta}^*, \boldsymbol{w}_t\right\rangle\right] + \frac{1}{2}h\kappa_1.
\end{aligned}$$

Since $\boldsymbol{w}_t$ is a zero-mean Gaussian vector, taking expectation over $\{\boldsymbol{w}_t\}_t$ on both sides removes the second term on the r.h.s. And following the same argument for (EC.24), we obtain that

$$\mathbb{E}_{\mathsf{SGP}}\left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right] \le \frac{\|\boldsymbol{\theta}^*\|_2^2}{2T\eta} + \frac{\eta L^2}{2} + \frac{\eta d\sigma^2}{2} + \frac{1}{2}h\kappa_1. \tag{EC.31}$$

Therefore, substituting (EC.29) and (EC.31) into (EC.26), and then plugging in $h = \frac{\eta\overline{K}B_x^2}{2}$ results in

$$\mathcal{R}(\mathsf{SGP}; \mathbb{P}) \le \frac{(1+T)\eta L^2}{n} + \frac{\|\boldsymbol{\theta}^*\|_2^2}{2T\eta} + \frac{\eta L^2}{2} + \frac{\eta d\sigma^2}{2} + \frac{\eta\kappa_1\overline{K}B_x^2}{4}. \tag{EC.32}$$

Lastly, taking $\eta^* = \sqrt{\frac{\|B_\theta\|_2^2}{2T} \Big/ \left(\frac{(1+T)L^2}{n} + \frac{L^2}{2} + \frac{d\sigma^2}{2} + \frac{\kappa_1\overline{K}B_x^2}{4}\right)}$ gives the desired upper bound:

$$\begin{aligned}
\sup_{\mathbb{P}\in(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{SGP}_h; \mathbb{P}) &\le \sqrt{2}B_\theta\sqrt{\frac{L^2}{n} + \frac{8dL^2\ln(1/\delta)}{n^2\varepsilon^2} + \frac{L^2 + \kappa_1\overline{K}B_x^2}{2n^2} + \frac{L}{n^3}} \\
&\le \sqrt{2}B_\theta B_x\bar{r}\sqrt{\frac{2}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2} + \frac{1+\kappa_1\overline{K}/\bar{r}^2}{2n^2}} \\
&\le \sqrt{2}B_\theta B_x\bar{r}\sqrt{\frac{4}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} \qquad \left(\text{since } n \ge \frac{1+\kappa_1\overline{K}/\bar{r}^2}{4}\right) \\
&= \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right),
\end{aligned}$$

where the sup operator is by noticing that the above analysis can be applied to any distributions.

□

EC.1.9. Proof of Theorem 7

Proof. The proof includes four steps. In the first three steps, we find some lower bounds for the minimax risk of an (ε,δ)-DP algorithm $\mathsf{A}:\mathcal{X}^n\times\mathcal{Y}^n\to\mathcal{C}$, where $\mathcal{C}:=\mathcal{B}(B_\theta)$. These lower bounds finally lead to a minimax risk of an (ε',δ')-DP d-dimensional classification problem. The sample complexity of DP classification problems for achieving a certain accuracy is well studied in DP literature, from which we can derive the accuracy under a specific sample size. Consequently, the accuracy provides an overall lower bound.

1) **Step 1: relax the feasible region of the** inf **problem & restrict the feasible region of the** sup **problem**

We first lower bound the minimax risk by relaxing the feasible region of the inf problem. Note that, in the minimax risk $\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}}\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})}\mathcal{R}(\mathsf{A};\mathbb{P})$, the inf is taken over $\mathcal{F}_{\varepsilon,\delta}$, a set of mappings from $\mathcal{X}^n\times\mathcal{Y}^n$ to a Euclidean ball $\mathcal{C}=\mathcal{B}(B_\theta):=\left\{\boldsymbol{\theta}\in\mathbb{R}^d:\|\boldsymbol{\theta}\|_2\le B_\theta\right\}$. If we relax the output space from an Euclidean ball $\mathcal{C}$ to a superset $\mathcal{C}_\infty:=\left\{\boldsymbol{\theta}\in\mathbb{R}^d:\|\boldsymbol{\theta}\|_\infty\le B_\theta\right\}$ a box set with $\|\boldsymbol{\theta}\|_\infty:=\max_{j=1,\dots,d}|\theta_j|$, the corresponding feasible set of mappings $\mathcal{F}^\infty_{\varepsilon,\delta}$, which includes all DP mappings from $\mathcal{X}^n\times\mathcal{Y}^n$ to $\mathcal{C}_\infty$, is therefore much larger. Thus, we get a lower bound by relaxing the feasible region of the inf problem:

$$\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}}\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})}\mathcal{R}(\mathsf{A};\mathbb{P})\ge\inf_{\mathsf{A}\in\mathcal{F}^\infty_{\varepsilon,\delta}}\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})}\mathcal{R}(\mathsf{A};\mathbb{P}). \tag{EC.33}$$

We further discover a lower bound by restricting the feasible region of the sup problem. Specifically, we restrict $\mathbb{P}$ to a smaller set of probability measures $\mathbb{P}_\mathcal{S}$ where $\mathcal{S}\in\mathcal{X}^n\times\mathcal{Y}^n$ is a n-samples dataset with a specific structure, and $\mathbb{P}_\mathcal{S}$ is a distribution generated from the given dataset $\mathcal{S}$ by assigning each sample probability $1/n$:

$$(\text{EC.33})\ge\inf_{\mathsf{A}\in\mathcal{F}^\infty_{\varepsilon,\delta}}\sup_{\mathbb{P}_\mathcal{S}}\mathcal{R}(\mathsf{A};\mathbb{P}_\mathcal{S}).$$

The inequality holds as $\mathcal{S}$ will be restricted to a specific structure to be elaborated later. As a result of that, $\mathbb{P}_\mathcal{S}$ is optimized over a more restrictive region, leading to a smaller objective value. Therefore, the r.h.s of the above inequality is a lower bound.

The dataset $\mathcal{S}$ is constructed as follows: let context matrix $\boldsymbol{X}:=\{\boldsymbol{x}_i\}_{i=1}^n$ be an $n\times d$ matrix, where $\boldsymbol{x}_i\in\mathcal{X}:=\{\frac{-B_x}{\sqrt{d}},\frac{B_x}{\sqrt{d}}\}^d$. Denote the average vector by $\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i$, and its sign vector by $sign(\bar{\boldsymbol{x}}):=(sign(\bar{x}_1),\dots,sign(\bar{x}_d))\in\{-1,1\}^d$. Then, the dataset $\mathcal{S}$ is constructed as

$$\mathcal{S}=(B_\theta\cdot\boldsymbol{X}sign(\bar{\boldsymbol{x}}),\boldsymbol{X}).$$

By our construction, dataset $\mathcal{S}$ depends only on the context matrix $\boldsymbol{X}$; therefore, the $\sup_{\mathbb{P}_\mathcal{S}}$ is actually optimizing over $\boldsymbol{X}$. We will use $\mathcal{S}(\boldsymbol{X})$ to indicate such dependency when necessary. We conclude this step by writing down the lower bound explicitly for later reference:

$$\begin{aligned}\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}}\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})}\mathcal{R}(\mathsf{A};\mathbb{P})&\ge\inf_{\mathsf{A}\in\mathcal{F}^\infty_{\varepsilon,\delta}}\sup_{\mathbb{P}_\mathcal{S}}\mathcal{R}(\mathsf{A};\mathbb{P}_\mathcal{S})\\&=\inf_{\mathsf{A}\in\mathcal{F}^\infty_{\varepsilon,\delta}}\sup_{\mathbb{P}_\mathcal{S}}\left\{\mathbb{E}_{\mathsf{A},\mathcal{D}\sim\mathbb{P}_\mathcal{S}^n}\left[\mathcal{L}(\mathsf{A}(\mathcal{D});\mathbb{P}_\mathcal{S})\right]-\min_{\boldsymbol{\theta}\in\mathcal{C}_\infty}\mathcal{L}(\boldsymbol{\theta};\mathbb{P}_\mathcal{S})\right\},\end{aligned} \tag{EC.34}$$

where $\mathcal{R}(\mathsf{A};\mathbb{P}_{\mathcal{S}})$ is the excess generalization risk of A under distribution $\mathbb{P}_{\mathcal{S}}$.

2) **Step 2: reduce excess generalization risk to excess empirical risk**

We notice that, since $\mathbb{P}_{\mathcal{S}}$ is an n-valued distribution generated from $\mathcal{S}$, for any $\boldsymbol{\theta}$, the population risk $\mathcal{L}(\boldsymbol{\theta};\mathbb{P}_{\mathcal{S}})$ w.r.t. $\mathbb{P}_{\mathcal{S}}$ is equal to an empirical risk $\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S})$ w.r.t. $\mathcal{S}$ by definitions of $\mathcal{L}$ and $\widehat{\mathcal{L}}$, i.e., $\mathcal{L}(\boldsymbol{\theta};\mathbb{P}_{\mathcal{S}}) = \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S})$. Therefore, (EC.34) becomes

$$(\text{EC.34}) = \inf_{\mathsf{A}\in\mathcal{F}^{\infty}_{\varepsilon,\delta}} \sup_{\mathbb{P}_{\mathcal{S}}} \left\{ \mathbb{E}_{\mathsf{A}} \left[\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n_{\mathcal{S}}} \left[\widehat{\mathcal{L}}(\mathsf{A}(\mathcal{D});\mathcal{S}) \right] \right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_{\infty}} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \right\}. \tag{EC.35}$$

The inner expectation in the first term is taken over $\mathcal{D}\sim\mathbb{P}^n_{\mathcal{S}}$; that is, we i.i.d. draw n samples with replacement from a given $\mathcal{S}$. Thus, we can treat "subsampling $\mathcal{D}$ from $\mathcal{S}$, then run $\mathsf{A}(\mathcal{D})$" as a new algorithm B, which takes $\mathcal{S}$ as the input and outputs an estimator $\mathsf{A}(\mathcal{D})$. By a mild revision of notations only, we have

$$(\text{EC.35}) = \inf_{\substack{\mathsf{B}\text{: subsampling, then run }\mathsf{A},\\ \text{where }\mathsf{A}\in\mathcal{F}^{\infty}_{\varepsilon,\delta}}} \sup_{\mathcal{S}} \left\{ \mathbb{E}_{\mathsf{B}} \left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S}) \right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_{\infty}} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \right\}. \tag{EC.36}$$

It would be helpful to check whether algorithm B is DP. Following the definition of DP, we consider two neighboring dataset $\mathcal{T} := ((\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_k,y_k),\cdots,(\boldsymbol{x}_n,y_n))$ and $\mathcal{T}' := ((\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}'_k,y'_k),\cdots,(\boldsymbol{x}_n,y_n))$ that differ in k-th sample only. Datasets $\mathcal{T}$ and $\mathcal{T}'$ here do not necessarily follow the specific structure in Step 1; instead, they are conceptual here for checking if B is DP only. Let the set $\mathcal{I}\in\{1,\ldots,n\}^n$ be an index set with indices from i.i.d. sampling with replacement, and let $\mathcal{T}(\mathcal{I})$ be the resulting dataset with index set $\mathcal{I}$. Denote the number of different samples between $\mathcal{T}(\mathcal{I})$ and $\mathcal{T}'(\mathcal{I})$ by $\Delta(\mathcal{I}) := |\mathcal{T}(\mathcal{I})\backslash\mathcal{T}'(\mathcal{I})|$. As $\mathcal{T}$ and $\mathcal{T}'$ are neighboring and differ in k-th sample only, the value $\Delta(\mathcal{I})$ follows an n-trial Binomial distribution with success probability $1/n$; thus, it should be small with high probability. Specifically, we should have

$$\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\geq z+1\right] = \Pr\left[\mathsf{Binomial}(n,1/n)\geq z+1\right] \leq \exp\left(-z^2/3\right), \quad \forall z>0,$$

where the inequality follows from multiplicative Chernoff upper tail bound. Equivalently, the above implies that $\Delta(\mathcal{I})\geq 3\sqrt{\ln(1/\gamma)}+1 := u$ with probability at most γ. Now, we are ready to check if B is DP by definition: for any subset $\mathcal{U}$ of the output space of $\mathsf{B}(\mathcal{T})$, we have

$$\begin{aligned}
\Pr_{\mathsf{B}}\left[\mathsf{B}(\mathcal{T})\in\mathcal{U}\right] &\leq \Pr_{\mathsf{B}|\mathcal{I}}\left[\mathsf{B}(\mathcal{T})\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+\gamma\\
&= \Pr_{\mathsf{A}|\mathcal{I}}\left[\mathsf{A}(\mathcal{T}(\mathcal{I}))\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+\gamma\\
&\leq \left(e^{\Delta(\mathcal{I})\cdot\varepsilon}\Pr_{\mathsf{A}|\mathcal{I}}\left[\mathsf{A}(\mathcal{T}'(\mathcal{I}))\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]+\Delta(\mathcal{I})\delta e^{\Delta(\mathcal{I})\varepsilon}\right)\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+\gamma\\
&\leq e^{u\varepsilon}\Pr_{\mathsf{B}|\mathcal{I}}\left[\mathsf{B}(\mathcal{T}')\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+(u\delta e^{u\varepsilon}+\gamma)\\
&\leq e^{\varepsilon'}\Pr_{\mathsf{B}}\left[\mathsf{B}(\mathcal{T}')\in\mathcal{U}\right]+\delta',
\end{aligned}$$

where the third line follows from the fact that A is (ε,δ)-DP and Group Privacy Lemma (Vadhan 2017, Lemma 2.2). Therefore, the algorithm B is (ε',δ')-DP with $\varepsilon' := u\varepsilon$ and $\delta' := u\delta e^{u\varepsilon} + \gamma$. Moreover, we notice that algorithm B is very restrictive in the sense that it must follow a "subsampling, then run A" framework. If we remove the framework requirement and only require $\mathsf{B} \in \mathcal{F}^{\infty}_{\varepsilon',\delta'}$ a set of (ε',δ')-DP mappings from $\mathcal{X}^n \times \mathcal{Y}^n$ to $\mathcal{C}_\infty$, we will get a lower bound to (EC.36):

$$\text{(EC.36)} \geq \inf_{\substack{\mathsf{B} \in \mathcal{F}^{\infty}_{\varepsilon',\delta'}\text{: subsampling, then run A,} \\ \text{where } \mathsf{A} \in \mathcal{F}^{\infty}_{\varepsilon,\delta}}} \sup_{\mathcal{S}} \left\{ \mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_\infty} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \right\}$$

$$\geq \inf_{\mathsf{B}\in\mathcal{F}^{\infty}_{\varepsilon',\delta'}} \sup_{\mathcal{S}} \left\{ \mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_\infty} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \right\}. \tag{EC.37}$$

3) **step 3: convert excess empirical risk to DP binary classification error**

Now, we start to analyze the excess empirical risk $\mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_\infty} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S})$ for any given dataset $\mathcal{S}$. Recall that the dataset $\mathcal{S}$ is constructed as $(\boldsymbol{X}\cdot sign(\bar{\boldsymbol{x}})\cdot B_\theta, \boldsymbol{X})$ in Step 1. Hence, the empirical minimizer is $\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathcal{C}_\infty} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) = sign(\bar{\boldsymbol{x}})\cdot B_\theta \in \mathcal{C}_\infty$, and the empirical risk $\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}};\mathcal{S}) = 0$. We therefore only need to focus our attention on $\mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right]$. It is straightforward to show

$$\begin{aligned}
\mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] &= \mathbb{E}_{\mathsf{B}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(r\cdot(y_i - \mathsf{B}(\mathcal{S})^\top \boldsymbol{x}_i)^+ + (1-r)\cdot(\mathsf{B}(\mathcal{S})^\top \boldsymbol{x}_i - y_i)^+\right)\right] \\
&\geq \min\{r,1-r\}\mathbb{E}_{\mathsf{B}}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\left(sign(\bar{\boldsymbol{x}})\cdot B_\theta - \mathsf{B}(\mathcal{S})\right)^\top \boldsymbol{x}_i\right|\right] \\
&\geq \min\{r,1-r\}\mathbb{E}_{\mathsf{B}}\left[\left|\left(sign(\bar{\boldsymbol{x}})\cdot B_\theta - \mathsf{B}(\mathcal{S})\right)^\top \bar{\boldsymbol{x}}\right|\right]. \qquad \text{(by triangular inequality)}
\end{aligned}$$

By simple algebra, the absolute value in the expectation operator is

$$\begin{aligned}
\left|\left(sign(\bar{\boldsymbol{x}})\cdot B_\theta - \mathsf{B}(\mathcal{S})\right)^\top \bar{\boldsymbol{x}}\right| &= \left|\sum_{j=1}^{d} |\bar{x}_j|\left(B_\theta - \mathsf{B}(\mathcal{S})_j \cdot sign(\bar{\boldsymbol{x}})_j\right)\right| \\
&= \sum_{j=1}^{d} |\bar{x}_j|\left(B_\theta - \mathsf{B}(\mathcal{S})_j \cdot sign(\bar{\boldsymbol{x}})_j\right) \qquad \text{(Since } |\mathsf{B}(\mathcal{S})_j| \leq B_\theta\text{)} \\
&\geq B_\theta \cdot \sum_{j=1}^{d} |\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq sign(\mathsf{B}(\mathcal{S})_j)\right\}.
\end{aligned}$$

Putting the above analysis together, we further obtain a lower bound:

$$\begin{aligned}
\text{(EC.37)} &\geq \min\{r,1-r\}B_\theta \cdot \inf_{\mathsf{B}\in\mathcal{F}^{\infty}_{\varepsilon',\delta'}} \sup_{\mathcal{S}} \mathbb{E}_{\mathsf{B}}\left[\sum_{j=1}^{d} |\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq sign(\mathsf{B}(\mathcal{S})_j)\right\}\right] \\
&= \min\{r,1-r\}B_\theta \cdot \inf_{\mathsf{B}\in\mathcal{F}^{\infty}_{\varepsilon',\delta'}} \sup_{\boldsymbol{X}\in\mathcal{X}^n} \mathbb{E}_{\mathsf{B}}\left[\sum_{j=1}^{d} |\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq sign(\mathsf{B}(\mathcal{S}(\boldsymbol{X}))_j)\right\}\right]
\end{aligned}$$

$$\geq \min\{r, 1-r\} B_\theta \cdot \inf_{\substack{\mathsf{C}: \mathcal{X}^n \to \{-1,1\}^d; \\ \mathsf{C} \text{ is } (\varepsilon', \delta')\text{-DP}}} \sup_{\boldsymbol{X} \in \mathcal{X}^n} \mathbb{E}_{\mathsf{C}} \left[\sum_{j=1}^{d} |\bar{x}_j| \, \mathbb{1}\left\{ sign(\bar{\boldsymbol{x}})_j \neq \mathsf{C}(\boldsymbol{X})_j \right\} \right], \tag{EC.38}$$

where the second line is due to the dependence between dataset $\mathcal{S}$ and feature matrix $\boldsymbol{X}$ that $\mathcal{S}$ does not provide more information than $\boldsymbol{X}$; the third line follows a similar idea to (EC.37), and $\mathsf{C}: \mathcal{X}^n \to \{-1,1\}^d$ is a DP binary classifier. The expectation term in (EC.38) can be viewed as the worst-case expected error of estimating the $sign$ vector of the (column-wise) average vector of $\boldsymbol{X}$ when using a DP binary classifier C.

4) **step 4: bounding the classification error with sample complexity**

To facilitate further analysis, we use $\mathsf{Error}(\mathsf{C}, \boldsymbol{X}) := \mathbb{E}_{\mathsf{C}} \left[\sum_{j=1}^{d} |\bar{x}_j| \, \mathbb{1}\left\{ sign(\bar{\boldsymbol{x}})_j \neq \mathsf{C}(\boldsymbol{X})_j \right\} \right]$ to denote the expected error of a DP binary classification algorithm C for a given $n \times d$ matrix $\boldsymbol{X}$. Denote the smallest number of samples to achieve a certain minimax risk $\alpha > 0$ as the sample complexity:

$$\mathsf{S}(\alpha, \varepsilon, \delta) := \min\{n : \inf_{\mathsf{C} \in \mathcal{F}_{\varepsilon,\delta}, \text{classifier } \mathsf{C}} \sup_{\boldsymbol{X} \in \mathcal{X}^n} \mathsf{Error}(\mathsf{C}, \boldsymbol{X}) \leq \alpha\}.$$

By Proposition 1 and Lemma D.2 in Asi et al. (2021), for ε', δ' defined in step 3, the sample complexity satisfies

$$\mathsf{S}(\alpha, \varepsilon', \delta') \geq \Omega\left(\frac{\sqrt{d}}{\alpha \varepsilon' \ln d} \right).$$

The lower bound on sample complexity implies a lower bound on the minimax risk:

$$\inf_{\mathsf{C} \in \mathcal{F}_{\varepsilon',\delta'}, \text{classifier } \mathsf{C}} \sup_{\boldsymbol{X} \in \mathcal{X}^n} \mathsf{Error}(\mathsf{C}, X) \geq \Omega\left(\frac{\sqrt{d}}{n\varepsilon' \ln d} \right) = \Omega\left(\frac{\sqrt{d}}{n\varepsilon \ln d} \right).$$

Combining the preceding four steps, we obtain

$$\inf_{\mathsf{A} \in \mathcal{F}_{\varepsilon,\delta}} \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \mathcal{R}(\mathsf{A}; \mathbb{P}) \geq \widetilde{\Omega}\left(\frac{\sqrt{d}}{n\varepsilon} \right), \tag{EC.39}$$

where a logarithmic factor $\ln d$ in the denominator of r.h.s of (EC.39) is hidden. Moreover, since the oracle complexity of stochastic convex optimization is $\Omega(\frac{1}{\sqrt{n}})$, the minimax risk is therefore further lower bounded by the maximum between oracle complexity and (EC.39):

$$\begin{aligned} \inf_{\mathsf{A} \in \mathcal{F}_{\varepsilon,\delta}} \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \mathcal{R}(\mathsf{A}; \mathbb{P}) &\geq \widetilde{\Omega}\left(\max\left(\frac{1}{\sqrt{n}}, \frac{\sqrt{d}}{n\varepsilon} \right) \right) \\ &= \widetilde{\Omega}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon} \right), \end{aligned}$$

where the second line gives the desired bound.

□

EC.1.10. Proposition EC.1 and Proof

Proposition EC.1 (Inconsistent Order Quantity by Output Perturbation). *Consider a sampling case DP-NV problem without features. Assume demand $y \sim \mathbb{P} \in \mathcal{P}([0,\bar{y}])$ where $\mathbb{P}$ has a CDF F supported on $[0,\bar{y}]$ with bounded $\bar{y} > 0$. Let optimal order quantity be $q^* := \inf_{y\in\mathbb{R}}\{y : F(y) \geq r\}$, and let non-private data-driven order quantity be $\widehat{q} := \inf_{y\in\mathbb{R}}\{y : \widehat{F}(y) \geq r\}$. For any given (ε,δ), let $\widehat{q}^{\mathcal{G}} := \widehat{q} + \mathcal{N}(0,\sigma^2)$ with proper σ by Gaussian Mechanism 1. Then, there exists a strictly positive constant $C > 0$ such that*

$$\Pr\left[\left|\widehat{q}^{\mathcal{G}} - q^*\right| \geq \epsilon\right] \geq C > 0, \quad \forall \epsilon \in (0, \max\{q^*, \bar{y} - q^*\}),$$

where the probability is taken over Gaussian noise. In other words, $\widehat{q}^{\mathcal{G}}$ is inconsistent.

Proof. For given demand observations $\boldsymbol{y} := \{y_i\}_{i=1}^n$ and critical ratio $r \in (0,1)$, we first notice that $\widehat{q} := \inf_{y\in\mathbb{R}}\left\{y : \widehat{F}(y) \geq r\right\} = \boldsymbol{y}_{(\lceil nr \rceil)}$, where $\boldsymbol{y}_{(\lceil nr \rceil)}$ is the $\lceil nr \rceil^{th}$-smallest value. As a result, the Global Sensitivity of $\widehat{q}$ is $\mathsf{GS}_{\widehat{q}} = \sup_{\boldsymbol{y}\sim\boldsymbol{y}'} |\widehat{q}(\boldsymbol{y}) - \widehat{q}(\boldsymbol{y}')| = \bar{y}$. This is achievable by a pair of neighboring demand samples with $\lceil nr \rceil - 1$ points at 0 and $n - \lceil nr \rceil$ points at $\bar{y}$, and differ only in the $\lceil nr \rceil^{th}$-smallest point. Whether the $\lceil nr \rceil^{th}$-smallest point is at 0 or $\bar{y}$ determines whether the order quantity is 0 or $\bar{y}$. Consequently, by Gaussian Mechanism 1, there exists a valid $\sigma \geq \sqrt{2\ln(1.25/\delta)} \cdot \bar{y}/\varepsilon$ such that $\widehat{q}^{\mathcal{G}} := \widehat{q} + \mathcal{N}(0,\sigma^2)$ is (ε,δ)-DP. Then for any given $\boldsymbol{y}$ and $\epsilon \in (0, \max\{q^*, \bar{y} - q^*\})$, we have

$$\begin{aligned}
\Pr\left[\left|\widehat{q}^{\mathcal{G}} - q^*\right| \geq \epsilon\right] &= \Pr\left[\left|\widehat{q} + \mathcal{N}(0,\sigma^2) - q^*\right| \geq \epsilon\right] \\
&= \Pr\left[\widehat{q} + \mathcal{N}(0,\sigma^2) - q^* \geq \epsilon\right] + \Pr\left[\widehat{q} + \mathcal{N}(0,\sigma^2) - q^* \leq -\epsilon\right] \\
&\geq \Pr\left[0 + \mathcal{N}(0,\sigma^2) - q^* \geq \epsilon\right] + \Pr\left[\bar{y} + \mathcal{N}(0,\sigma^2) - q^* \leq -\epsilon\right] \\
&= \Pr\left[\mathcal{N}(0,\sigma^2) \geq q^* + \epsilon\right] + \Pr\left[\mathcal{N}(0,\sigma^2) \geq \bar{y} + \epsilon - q^*\right], \qquad \text{(EC.40)}
\end{aligned}$$

where all probabilities are taken with respect to the algorithm's randomness, i.e. the Gaussian noise. Since the standard Gaussian $z \sim \mathcal{N}(0,1^2)$ admits a tail bound that $\Pr[z \geq t] \geq \frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} \cdot e^{-t^2/2}, \forall t > 0$, it naturally leads to a lower bound for (EC.40):

$$\text{(EC.40)} \geq \frac{1}{\sqrt{2\pi}} \left(\frac{t_1}{t_1^2+1} \cdot e^{-t_1^2/2} + \frac{t_2}{t_2^2+1} \cdot e^{-t_2^2/2} \right) > 0,$$

with $t_1 := (q^* + \epsilon)/\sigma$ and $t_2 := (\bar{y} + \epsilon - q^*)/\sigma$.

□

EC.2. Omitted Materials for Sections 2 and 3

EC.2.1. Compare to the Concurrent Work Zhao et al. (2025)

Table EC.2, furnished with more details compared to Table 4 in the main text, summarizes the technical setup of Zhao et al. (2025) and our work. Because their algorithm is a gradient-based algorithm, comparisons here are mainly about gradient-based algorithms.

Table EC.2 Comparison with the Concurrent Work Zhao et al. (2025)

	Assumptions on Groundtruth Demand Model			Privacy	Analysis Tool	Perf. Metric
	Structure	Features	Error Term			
Zhao et al. (2025)	$y = \langle \boldsymbol{\theta}^\star, \boldsymbol{x} \rangle + \epsilon(\boldsymbol{x})$	$\cdot\ x_1 \equiv 1$ $\cdot\ \boldsymbol{x}_{-1}$ SubGaussian	$\forall \boldsymbol{x}$: $\cdot$ CDF $F^{-1}_{\epsilon\mid\boldsymbol{x}}(r) = 0$ $\cdot$ pdf $f_{\epsilon\mid\boldsymbol{x}}$ exists $\cdot f_{\epsilon\mid\boldsymbol{x}}$ Lipschitz cts. $\cdot \exists f_u > 0$ s.t. $f_{\epsilon\mid\boldsymbol{x}} \le f_u$ $\cdot$ more assumptions‡	Gaussian-DP	High-dim. stats.	Est. error
This work	$y = \langle \boldsymbol{\theta}^\star, \boldsymbol{x} \rangle + \epsilon^\star(\boldsymbol{x})$	$\boldsymbol{x}$ bounded	No assumptions	(ε, δ)-DP	Convex opt.	Generalization Risk

Note. Comparison of assumptions in this table is stated for gradient-based policies (Section 3.4), because the only algorithm in Zhao et al. (2025) is gradient-based. The demand model in the concurrent work is stated with an (unknown) population-level minimizer $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[r\cdot(y-\langle\boldsymbol{\theta},\boldsymbol{x}\rangle)^+ + (1-r)(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle - y)^+\right]$ and a so-called "observational error term" $\epsilon(\boldsymbol{x}) := y - \langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle$; so any assumptions imposed on their error term are unverifiable in practice. ‡More (unverifiable) assumptions: (i) $\exists f_l > 0$ such that $\inf_{t\in[0,1],\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} \mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon\mid\boldsymbol{x}}(t\langle\boldsymbol{w},\boldsymbol{v}\rangle)\langle\boldsymbol{w},\boldsymbol{v}\rangle^2\right] \ge f_l$ where $\boldsymbol{w} := \Sigma^{-1/2}\boldsymbol{x}$ and $\Sigma := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$; and (ii) $\exists f_l' > 0$ such that $\inf_{|u|\le 1} \frac{1}{2u}\int_{-u}^{u} f_{\epsilon\mid\boldsymbol{x}}(v)\,dv \ge f_l'$. Abbreviations: "Perf. Metric"= performance metric; "Est.error"= estimation error; "High-dim. stats."= high-dimensional statistics; "Convex opt."= convex optimization; "cts."= continuous.

Demand model. Our model is the most natural and straightforward model $y = \langle\boldsymbol{\theta}^\star, \boldsymbol{x}\rangle + \epsilon^\star(\boldsymbol{x})$ without assumptions except linearity and boundedness. In contrast, the demand model in Zhao et al. (2025) is rather complicated: it is stated with the optimal population-level minimizer

$$\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[r\cdot(y-\langle\boldsymbol{\theta},\boldsymbol{x}\rangle)^+ + (1-r)(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle - y)^+\right],$$

and a so-called "observational error term" $\epsilon(\boldsymbol{x}) := y - \langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle$. Please pay attention to $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^\star$, they are different in general. Additionally, they imposed many restrictive assumptions on the observational error term, which appear to create a backdoor to bypass the inherent hardness of the private newsvendor problem; we will provide more details later when discussing convergence rates. Moreover, these assumptions are unverifiable in practice, because one has to figure out the optimal $\boldsymbol{\theta}^*$ first. And only after that, the observational error term can be properly defined. Lastly, these assumptions implicitly force demand y to be continuous at least in some regions, which might conflict with business applications where demand is discrete and products are indivisible. In contrast, we do NOT make any assumptions on the error term.

Convergence rates. We first repeat the convergence rates:

$$\text{(ours)}\quad \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right); \qquad \text{(theirs)}\quad \widetilde{\mathcal{O}}\left(\frac{f_u}{f_l^2}\frac{d}{n} + \frac{1}{f_u}\left(\frac{d+\ln n}{n\varepsilon}\right)^2\right).$$

It is spurious that their rate has a better dependence on sample size n. But this is because of the technical assumptions they imposed on their demand model, rather than solely by a more involved analysis.

Assumptions made by Zhao et al. (2025), especially the two we restate in the note of Table 4, restrict their observational error term $\epsilon(\boldsymbol{x})$ to stay around the origin of loss functions with reasonable probability. Because around the origin, the smoothed loss function locally behaves like a strongly convex function (see

Figure 3a), they can employ strong convexity and get a faster convergence rate matching that of DP-SCO under strongly convex functions. If their assumptions do not hold, i.e., $f_l = 0$, then their convergence rate becomes meaningless. Therefore, their convergence rates and upper bounds should be understood as bounds for distributions in a smaller set satisfying their strong assumptions. To better support this statement, one fact from our analysis of lower bounds is that the lower bound is achieved by a set of "discrete" distributions that stay away from the origin. That means their assumptions rule out the distributions that inherently capture the hardness of the private newsvendor, allowing them to break through the information-theoretical lower bound (Theorem 7) on general nonsmooth DP-SCO for a faster rate. In summary, their faster rates are consequences of assumptions imposed on the demand model, rather than the consequence of leveraging the structure of the newsvendor problem as they claimed in their abstract.

Questionable optimality claim. The authors of Zhao et al. (2025) claim their parameter convergence rate is optimal under their setting. However, their optimality claim is not supported with proper evidence of a minimax lower bound. The minimax lower bound invoked by the authors is Theorem 4.1 in Cai et al. (2021). But in fact Cai et al. (2021, Theorem 4.1) proves the minimax lower bound for a different linear model: a model with a zero-mean independent Gaussian error term. Besides this obvious difference, there are many other nuances that make invoking Cai et al. (2021, Theorem 4.1) improper. First, the zero-mean Gaussian's r-th quantile is not 0 (unless $r = 1/2$). But Zhao et al. (2025) assumes the r-th quantile of their observational error term to be 0, see their assumption $F^{-1}_{\epsilon|\boldsymbol{x}}(r) = 0, \forall \boldsymbol{x}$; thus zero-mean Gaussian is not a valid "observational error term". As a result, the invoked lower bound is not applicable. Second, the lower bound in Cai et al. (2021, Theorem 4.1) is proved by tracing attack argument (now known as score attack Cai et al. 2023), which heavily relies on the zero-mean nature and the independence of the Gaussian error term, see the second paragraph in the proof of supporting technical Lemma 4.1 in Cai et al. (2021). It is unclear immediately whether the analysis can still go through, if the error term is not zero-mean and not independent. Third, other assumptions imposed on the observational error term by Zhao et al. (2025) further obfuscate the connection with Cai et al. (2021), making it unclear whether zero-mean Gaussian satisfies these assumptions. For these three reasons, invoking Cai et al. (2021) is improper, and there is a necessity to formally prove a minimax lower bound under their setting. Unfortunately, the authors of Zhao et al. (2025) did not provide any proof or make the connection with Cai et al. (2021) clear. They simply claimed optimality and highlighted it as a main contribution. In contrast, we clearly noticed the discrepancy and thus provide a formal proof of the lower bound; see Theorem 7.

Minor distinctions. Other minor distinctions are summarized below. (i) Practicability. Our algorithms are implementable. In contrast, either their Algorithm 1 or update rule Eq.(7) requires information of $\Sigma := \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]$ or f_l and f_u to run with provably theoretical guarantees. (ii) Sample complexity. Our bounds of gradient-based policies hold for $n \geq 2$, whereas bounds in Zhao et al. (2025) hold only for $n \gtrsim \sqrt{T} \cdot \frac{d+\ln n}{\varepsilon f_l}$ (see Theorem 2 in their work), an incalculable value if no available information of the observational error term. Again, if their assumptions do not hold, i.e., $f_l = 0$, the sample complexity is meaningless.

Contributions. Besides the above distinctions, our work has more contributions. (i) We propose two more algorithms OP and SGP, where OP follows a completely different design philosophy and requires new analytics tools. (ii) We also fix a technical bug of privacy accounting that appears in the first paper on Objective Perturbation (Kifer et al. 2012); see Lemma EC.1 in the Appendix for details. Building upon derived performance bounds, (iii) we further examine the internal impact of DP in Section 4; and (iv) investigate its external impact, i.e., supply chain implications, in Section 5.

EC.2.2. Comparison between Convolution Smoothing and Moreau Envelope

The problem considered in our work technically belongs to nonsmooth DP-SCO. In nonsmooth DP-SCO literature, the most common smoothing technique is the (Standard) Moreau Envelope. Because we chose convolution smoothing for the studied problem, we would like to give a comparison between convolution and Moreau Envelope.

EC.2.2.1. Introduction to Moreau Envelope

Moreau Envelope is a smoothing technique that returns a smooth approximation function $c_{\mathsf{ME}}(\boldsymbol{u})$ of the original nonsmooth cost function $c:\mathbb{R}^m\to\mathbb{R}$. The approximation function $c_{\mathsf{ME}}(\boldsymbol{u})$ is given by a minimization problem,

$$\text{(Moreau Envelope)}\quad c_{\mathsf{ME}}(\boldsymbol{u})=\inf_{\boldsymbol{x}\in\mathbb{R}^m}\{c(\boldsymbol{x})+\frac{1}{2h}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2\},\quad \forall \boldsymbol{u}\in\mathbb{R}^m. \tag{EC.41}$$

When $m=1$ and $c(u)=ru^+ +(1-r)(-u)^+$, its Moreau approximation $c_{\mathsf{ME}}(u)$ has three pieces and admits a closed-form expression:

$$c_{\mathsf{ME}}(u)=\begin{cases}(r-1)u-h\frac{(1-r)^2}{2}, & \text{if } u\le (r-1)h;\\ \frac{u^2}{2h}, & \text{if } (r-1)h<u<rh;\\ ru-h\frac{r^2}{2}, & \text{if } u\ge rh.\end{cases}$$

The three pieces are plotted in Figure EC.3, compared to the newsvendor loss function in black. Formally, the smoothed function c_{ME} by Moreau has many similar properties to that by convolution.

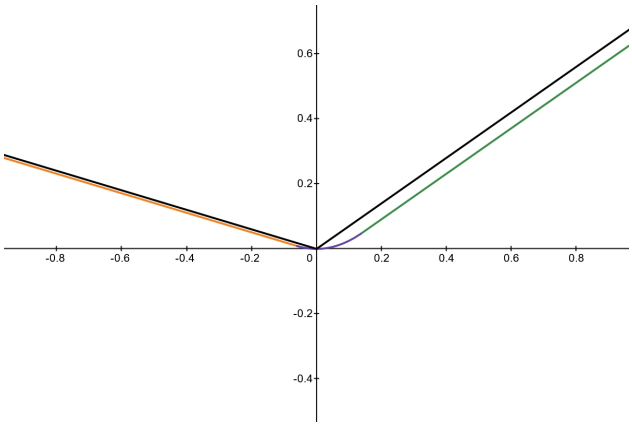
Lemma EC.4 (Properties of c_{ME}, Beck and Teboulle (2012)). *Let $c:\mathbb{R}\to\mathbb{R}$ be the newsvendor loss function, which is $\bar{r}=\max\{r,1-r\}$-Lipschitz continuous. Then its smoothed function by Moreau Envelope c_{ME} possesses following properties:*

1. *c_{ME} is first- and second-order differentiable;*
2. *c_{ME} is convex, $\bar{r}$-Lipschitz continuous, and $1/h$-smooth;*
3. *The approximation error is*

$$-h\cdot\frac{\max\{r^2,(1-r)^2\}}{2}\le c_{ME}(u)-c(u)\le 0,\quad \forall u\in\mathbb{R}.$$

It is evident that these properties are analogous to properties of convolution in Lemma 5. Therefore, in terms of approximation, Moreau and convolution are very similar.

Figure EC.3 The smoothed function by Moreau has three pieces highlighted in orange, purple, and green. The black curve is the newsvendor loss function.



EC.2.2.2. Technical Challenges

However, there are many technical challenges brought about by Moreau.

1. First, the approximation error by Moreau is noticeably higher than that by convolution. Specifically, the error by Moreau is linear in h, while the error by convolution might be faster than h, which depends on the kernel function chosen. If the kernel function has a light tail with an exponential decay rate, then the approximation error by convolution can be exponentially small. The poor approximation of Moreau poses a significant challenge: gaining smoothness from Moreau is more costly than from convolution. Recall that, we hope to gain smoothness to apply algorithmic stability lemma, and it is natural to have smoothness at a lower price; otherwise, the approximation error may harm convergence rates. While it may need further investigation to determine whether the linear rate error by Moreau harms convergence rates, the exponentially small error by convolution is definitely preferred.
2. It is evident from Figure EC.3 that c_{ME} approximates from below. But recall that the newly developed risk decomposition (7) heavily relies on the fact that convolution approximates from above. The fact that Moreau approximates from below means the new decomposition is not directly applicable, and all analyses developed in this work may not hold for algorithms using Moreau.

Given these two challenges here, more technical tools are deemed necessary to uncover the potential of Moreau. As we have developed tools to take advantage of convolution, we would like to leave the exploration of Moreau for future research.

EC.2.2.3. Generalized Moreau Envelope Does Not Help

The form of (EC.41) is typically known as the Standard Moreau Envelope. There is a more generalized Moreau Envelope that uses functions other than $\frac{1}{2}\|\cdot\|_2^2$. It is natural to wonder whether the previously mentioned challenges can be addressed by the generalized Moreau Envelope. Unfortunately, the answer is negative.

We begin the discussion with an introduction to the Generalized Moreau Envelope. Let $\phi:\mathbb{R}^m\to\mathbb{R}$ be a β-smooth function satisfying some regularity conditions (Condition 1).

Condition 1. *The smooth function $\phi:\mathbb{R}^m\to\mathbb{R}$ should satisfy following conditions:*

1. *function ϕ is continuously differentiable, strictly convex, and β-smooth;*
2. *its Fenchel conjugate function $\phi^\star(\boldsymbol{y}):=\sup_{\boldsymbol{u}\in dom\,\phi}\{\langle\boldsymbol{y},\boldsymbol{u}\rangle-\phi(\boldsymbol{u})\}$ exists, and the domain of $\phi^\star$ is a superset of subgradients of g, i.e., $dom\,\phi^\star\supseteq\cup_{\boldsymbol{x}\in\mathcal{X}}\partial c(\boldsymbol{x})$.*

There are many choices of function ϕ as listed in Table EC.3. For any ϕ, let $\phi_h(\cdot):=h\phi(\cdot/h)$. Then, the

Table EC.3 Eligible Functions $\phi(\cdot)$ for (Generalized) Moreau Envelope

Functions	$\phi(\boldsymbol{x})$	dom ϕ	Smoothness†	$\phi^\star(\boldsymbol{y})$	dom $\phi^\star$	Convexity†
Energy‡	$\frac{1}{p}\lvert x\rvert^p$	$\mathbb{R}$	1	$\frac{1}{q}\lvert y\rvert^q$	$\mathbb{R}$	1
ℓ_2-norm	$\frac{\beta}{2}\lVert\boldsymbol{x}\rVert_2^2$	$\mathbb{R}^d$	β	$\frac{1}{2\beta}\lVert\boldsymbol{y}\rVert_2^2$	$\mathbb{R}^d$	$\frac{1}{\beta}$
Huber	$\begin{cases}\beta\lVert\boldsymbol{x}\rVert_2^2/2, & \lVert\boldsymbol{x}\rVert_2\le 1,\\ \beta\lVert\boldsymbol{x}\rVert_2-\beta/2, & \text{otherwise.}\end{cases}$	$\mathbb{R}^d$	β	$\frac{1}{2\beta}\lVert\boldsymbol{y}\rVert_2^2$	$\mathcal{B}(\beta)$	$\frac{1}{\beta}$
Hellinger	$\beta\sqrt{1+\lVert\boldsymbol{x}\rVert_2^2}$	$\mathbb{R}^d$	β	$-\sqrt{\beta^2-\lVert\boldsymbol{y}\rVert_2^2}$	$\mathcal{B}(\beta)$	$\frac{1}{\beta}$

† with respect to l_2-norm $\lVert\cdot\rVert_2$; ‡ $p,q>0$ s.t. $1/p+1/q=1$;

approximation function by generalized Moreau Envelope $c_{\mathsf{GME}}(\boldsymbol{x})$ is formally defined as

$$c_{\mathsf{GME}}(\boldsymbol{u}):=\inf_{\boldsymbol{x}\in\mathbb{R}^m}\{c(\boldsymbol{x})+\phi_h(\boldsymbol{u}-\boldsymbol{x})\},\quad\forall\boldsymbol{u}\in\mathbb{R}^m. \tag{EC.42}$$

Similarly, c_{GME} has many desired properties.

Lemma EC.5 (Properties of c_{GME}, Beck and Teboulle 2012). *Let $c:\mathbb{R}^m\to\mathbb{R}$ be a closed, proper, convex, and L-Lipschitz continuous function (w.r.t. $\lVert\cdot\rVert_2$), and let $\phi:\mathbb{R}^m\to\mathbb{R}$ be a β-smooth function satisfying Condition 1. Then the Generalized Moreau Envelope c_{GME} possesses following properties:*

1. *$\nabla c_{GME}(\boldsymbol{u})=\nabla\phi_h(\boldsymbol{u}-\boldsymbol{x}^*(\boldsymbol{u}))$, where $\boldsymbol{x}^*(\boldsymbol{u})$ is the minimizer to the r.h.s problem of Eq.(EC.42);*
2. *c_{GME} is convex, L-Lipschitz, and (β/h)-smooth, w.r.t. $\lVert\cdot\rVert_2$;*
3. *Let $\phi^\star:\mathbb{R}^m\to\mathbb{R}$ be the Fenchel conjugate of ϕ, and let $\overline{\phi^\star}:=\sup_{\boldsymbol{y}\in\mathcal{B}(L)}\phi^\star(\boldsymbol{y})$ be its uniform upper bound. Then, the approximation error is*

$$-h\overline{\phi^\star}\le c_{GME}(\boldsymbol{u})-c(\boldsymbol{u})\le h\phi(\boldsymbol{0}).$$

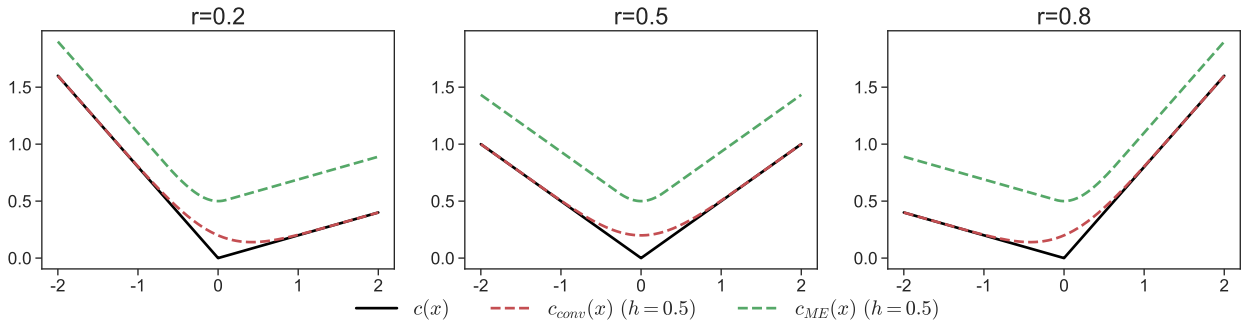
It is evident that the lemma here is an analogue to Lemmas 5 and EC.4. However, technical challenges brought about by Moreau are still there:

1. the approximation error is linear in h;
2. it is unclear immediately if Moreau can approximate from above.

Fortunately, the second issue can be resolved by choosing the Hellinger function as ϕ. If so, Table EC.3 suggests $\phi(\boldsymbol{0})=\beta$ and $\overline{\phi^\star}=0$, implying that Moreau Envelope now indeed approximates from above. But the approximation quality is much lower than convolution, see Figure EC.4. Therefore, while generalized

Moreau can approximate from above, the approximation quality is still lower than convolution. Because of these observations, we prefer convolution over Moreau.

Figure EC.4 Comparison between Convolution and Moreau. The ϕ function for Moreau is Hellinger.



EC.2.3. Extend OP to Other Nonsmooth Loss Functions

Extending OP to other nonsmooth functions is beyond the scope of our study. But for readers who are interested in the advantages of convolution in more general cases, we provide some preliminary results here.

Let $\boldsymbol{z} := (\boldsymbol{x}, y) \sim \mathbb{P}$ be the data point, where $\boldsymbol{x} \in \mathbb{R}^d$ is the feature vector, and $y \in \mathbb{R}$ is the label. We consider a nonsmooth loss function that follows the form

$$\ell(\boldsymbol{\theta}; \boldsymbol{z}) := \|A(\boldsymbol{z})\boldsymbol{\theta}\|_1 + f(\boldsymbol{z}^\top \boldsymbol{\theta}),$$

where $A : \mathbb{R}^{1+d} \to \mathbb{R}^m$ and $f : \mathbb{R} \to \mathbb{R}$ are known mappings; in additional, assume f is smooth in its argument, and thus the nonsmoothness issue purely comes from $\|\cdot\|_1$. This form covers many interesting cases; specially, it admits the newsvendor loss function as a special case. Below are some examples:

1. the newsvendor loss function. To see this, we can let $\boldsymbol{\theta} = (1, \boldsymbol{\theta}_{-1})$, $f : x \mapsto x/2$, and $A(\boldsymbol{z}) := (y, -\boldsymbol{x})/2$.
2. ReLU activation function $\max\{0, u\} = (|u| + u)/2$. It admits a reformulation with $A(\boldsymbol{z}) := \boldsymbol{z}^\top/2$ and $f(u) := u/2$, where u is the residual derived from $\boldsymbol{z}^\top\boldsymbol{\theta}$.
3. τ-soft-thresholding $(|u| - \tau)^+ = \left\| \begin{smallmatrix} (u+\tau)/2 \\ (u-\tau)/2 \end{smallmatrix} \right\|_1 - \tau$.

The considered nonsmooth structure contains a symmetric term $\|\cdot\|_1$. Therefore, we can apply convolution to this part. Denote $\boldsymbol{u} := A(\boldsymbol{z})\boldsymbol{\theta}$, we obtain a smooth approximation:

$$c_h(\boldsymbol{u}) := \int_{\boldsymbol{v} \in \mathbb{R}^m} c(\boldsymbol{u} + h\boldsymbol{v}) K(\boldsymbol{v})\, d\boldsymbol{v}\,.$$

It can be found that $c_h(\cdot)$ has similar properties to that in Lemma 5.

Lemma EC.6 (Properties of c_h). *Let $K(\boldsymbol{v}) = e^{-\frac{\|\boldsymbol{v}\|_2^2}{2}}/(\sqrt{2\pi})^m$ be the high-dimensional Gaussian kernel. Then, the approximation function c_h has the following properties:*

1. c_h *is convex,* $\sqrt{m}$*-Lipschitz and* $\sqrt{m}/h$*-smooth w.r.t.* $\|\cdot\|_2$*;*
2. $c_h(\boldsymbol{u}) = \|\boldsymbol{u}\|_1 + \sum_{j=1}^m \left(h\sqrt{2/\pi}\exp\left(-\frac{u_j^2}{2h^2}\right) - 2\,|u_j|\,\Phi\left(-\frac{|u_j|}{h}\right)\right)$.

Proof of Lemma EC.6. The first property is from Duchi et al. (2012). The second property is by direct calculation:

$$\begin{aligned} c_h(\boldsymbol{u}) = \int_{\boldsymbol{v}\in\mathbb{R}^m} \|\boldsymbol{u}+h\boldsymbol{v}\|_1 K(\boldsymbol{v})\,d\boldsymbol{v} &= \int_{v_m}\cdots\int_{v_1} \|\boldsymbol{x}+h\boldsymbol{v}\|_1 \Pi_{j=1}^m \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v_j^2}{2}\right)\right) dv_1\ldots dv_m \\ &= \|\boldsymbol{u}\|_1 + \sum_{j=1}^m \left(h\sqrt{2/\pi}\exp\left(-\frac{u_j^2}{2h^2}\right) - 2\,|u_j|\,\Phi\left(-\frac{|u_j|}{h}\right)\right). \end{aligned}$$

□

Immediately, we can upper bound the approximation gap as

$$\begin{aligned} c_h(\boldsymbol{u}) - c(\boldsymbol{u}) &= \sum_{j=1}^m \left(h\sqrt{2/\pi}\exp\left(-\frac{u_j^2}{2h^2}\right) - 2\,|u_j|\,\Phi\left(-\frac{|u_j|}{h}\right)\right) \\ &\le \sqrt{2/\pi}\cdot h\sum_{j=1}^m \exp\left(-\frac{u_j^2}{2h^2}\right), \quad \forall \boldsymbol{u}\in\mathbb{R}^m. \end{aligned}$$

The approximation gap is again exponentially small. We thus expect the analysis of newsvendor loss functions can be generalized to other nonsmooth functions with a term $\|\cdot\|_1$. Nevertheless, there is an additional factor of m, and the Lipschitz continuity and smoothness parameters are now also contingent on m. Further investigation is needed to determine whether the dimension dependency affects convergence rates. We leave this interesting question for exploration in the future.

EC.2.4. Extend OP to Newsvendor Problems with a (restricted) Discrete Demand Model

In Theorem 2 (performance of OP), we assume the error term to be Lipschitz continuous. However, this assumption may (or may not) implicitly imply a continuous demand model. Here, we consider a restricted discrete demand model and show that OP still works under mild assumptions.

For technical reasons, in addition to Assumption 1, we impose more assumptions on the demand model.

Assumption EC.1 ((Restricted) Discrete Demand Model). *We assume*

1. *(Exogeneity) the error term* $\epsilon^\star$ *is exogenous;*
2. *(Intercept) the first element of feature vector is* $x_1 \equiv 1$*;*
3. *(Local Behavior) there exists a threshold* $\tau > 0$ *such that* $\Pr_{\epsilon^\star}\left[\left|\epsilon^\star - F_{\epsilon^\star}^{-1}(r)\right| \le t\right] \le \exp\left(-1/t^2\right), \forall t \le \tau$*. The CDF of* $\epsilon^\star$ *is continuous at its* r*-th quantile.*

The first two assumptions are common in the literature, and are without loss of generality. The third assumption is mild, because it only imposes restrictions on $\epsilon^\star$'s local behavior around its r-th quantile $F_{\epsilon^\star}^{-1}(r)$. Outside the τ-neighborhood of $F_{\epsilon^\star}^{-1}(r)$, $\epsilon^\star$ is not required to follow any assumptions. Therefore, under this new assumption, $\epsilon^\star$ can be discrete outside the local region; thus demand y is also discrete at most points. We call it a restricted discrete demand model because it indeed has some restrictions around $F_{\epsilon^\star}^{-1}(r)$.

Moreover, we want to mention that the third part of Assumption EC.1 is unverifiable in practice. Therefore, the derived theorem below is just for illustrative purposes.

Theorem EC.1 (OP for Restricted Discrete Demand). *Suppose Assumptions 1 and EC.1 hold. If the kernel function is Gaussian kernel, $\delta \le e^{-\varepsilon/2}$, and $\varepsilon \ge \frac{\overline{K} B_x B_\theta}{\bar{r}\sqrt{n}} \cdot \max\left\{\frac{1}{\sqrt{2}\tau^2}, \frac{1}{\sqrt{\pi}}, \ln\left(\frac{\sqrt{n}}{2\sqrt{2}B_x B_\theta \bar{r}}\right)/\sqrt{2}\right\}$, then running Algorithm* OP *with $\lambda = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}}$ and $h = \frac{\overline{K} B_x^2}{\lambda n \varepsilon}$ yields*

$$\sup_{\mathbb{P}} \mathcal{R}(\mathsf{OP}; \mathbb{P}) \le 4\sqrt{2} B_x B_\theta \bar{r} \sqrt{\frac{1}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} = \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right).$$

Proof of Theorem EC.1. From (EC.16) and (EC.17) and letting $\lambda = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}}$ and $h = \frac{\overline{K} B_x^2}{\lambda n \varepsilon}$, we know

$$\mathcal{R}(\mathsf{OP}; \mathbb{P}) \le 2B_\theta \sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}} + [\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*)].$$

Since $\sigma^2 = L^2(8\ln(1/\delta) + 4\varepsilon)/\varepsilon^2$, and $\delta \le e^{-\varepsilon/2}$, the first term in the preceding inequality can be simplified; thus

$$\mathcal{R}(\mathsf{OP}; \mathbb{P}) \le 2\sqrt{2} B_x B_\theta \bar{r} \sqrt{\frac{1}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} + [\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*)]. \tag{EC.43}$$

By the fact $\ell_h(\boldsymbol{\theta}; \boldsymbol{x}, y) - \ell(\boldsymbol{\theta}; \boldsymbol{x}, y) \le \frac{h}{2}\int_{v:|v|\ge|y-\langle\boldsymbol{\theta},\boldsymbol{x}\rangle|/h} |v| K(v)\, dv$ and the fact kernel K is Gaussian, the approximation error in the bracket can be upper bounded as

$$\begin{aligned}
\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) &\le \frac{h}{2}\mathbb{E}_{\boldsymbol{x},y}\left[\int_{v:|v|\ge|y-\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle|/h} |v| K(v)\, dv\right] \\
&\le \frac{h}{2}\mathbb{E}_{\boldsymbol{x},y}\left[\sqrt{2/\pi}\exp\left(-\frac{(y-\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle)^2}{2h^2}\right)\right].
\end{aligned}$$

By Lemma EC.8, $\boldsymbol{\theta}^* := \boldsymbol{\theta}^\star + [F_{\epsilon^\star}^{-1}(r), 0, \ldots, 0]$. So,

$$y - \langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle = \langle\boldsymbol{\theta}^\star, \boldsymbol{x}\rangle + \epsilon^\star - \langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle = \epsilon^\star - F_{\epsilon^\star}^{-1}(r).$$

Therefore,

$$\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) \le \sqrt{1/(2\pi)} \cdot h \cdot \mathbb{E}_{\epsilon^\star}\left[\exp\left(-\frac{(\epsilon^\star - F_{\epsilon^\star}^{-1}(r))^2}{2h^2}\right)\right].$$

If parameters are properly chosen such that $\sqrt{h} \le \tau$, by the third part of Assumption EC.1, we have

$$\begin{aligned}
\mathbb{E}_{\epsilon^\star}\left[\exp\left(-\frac{(\epsilon^\star - F_{\epsilon^\star}^{-1}(r))^2}{2h^2}\right)\right] &= \mathbb{E}_{\epsilon^\star}\left[\exp\left(-\frac{(\epsilon^\star - F_{\epsilon^\star}^{-1}(r))^2}{2h^2}\right) \;\middle|\; |\epsilon^\star - F_{\epsilon^\star}^{-1}(r)| > \sqrt{h}\right] \cdot \mathbb{P}_{\epsilon^\star}\left[|\epsilon^\star - F_{\epsilon^\star}^{-1}(r)| > \sqrt{h}\right] \\
&\quad + \mathbb{E}_{\epsilon^\star}\left[\exp\left(-\frac{(\epsilon^\star - F_{\epsilon^\star}^{-1}(r))^2}{2h^2}\right) \;\middle|\; |\epsilon^\star - F_{\epsilon^\star}^{-1}(r)| \le \sqrt{h}\right] \cdot \mathbb{P}_{\epsilon^\star}\left[|\epsilon^\star - F_{\epsilon^\star}^{-1}(r)| \le \sqrt{h}\right] \\
&\le \exp\left(-\frac{1}{2h}\right) \cdot 1 + 1 \cdot \exp\left(-\frac{1}{h}\right) \\
&\le 2\exp(-1/h).
\end{aligned}$$

It immediately implies an upper bound on $\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*)$ as follows:

$$\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \sqrt{2/\pi} \cdot h \exp(-1/h). \tag{EC.44}$$

Plugging it back to the regret upper bound, we get

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq 2\sqrt{2} B_x B_\theta \bar{r} \sqrt{\frac{1}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} + \sqrt{2/\pi} \cdot h \exp(-1/h). \tag{EC.45}$$

To ensure the first term in (EC.45) dominates its second term, a sufficient condition is

$$\begin{cases} h \leq \tau^2 \\ \sqrt{2/\pi} h \leq 1 \\ \exp(-1/h) \leq 2\sqrt{2} B_x B_\theta \bar{r} \sqrt{\frac{1}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} \end{cases} \Leftarrow \begin{cases} \varepsilon \geq \frac{\overline{K} B_x B_\theta}{\sqrt{2}\bar{r}\tau^2} \cdot \frac{1}{\sqrt{n}} \\ \varepsilon \geq \frac{\overline{K} B_x B_\theta}{\bar{r}\sqrt{\pi}} \cdot \frac{1}{\sqrt{n}} \\ \varepsilon \geq \frac{\overline{K} B_x B_\theta}{\sqrt{2}\bar{r}} \cdot \frac{\ln\left(\frac{\sqrt{n}}{2\sqrt{2} B_x B_\theta \bar{r}}\right)}{\sqrt{n}} \end{cases}.$$

Therefore, if $\varepsilon \geq \frac{\overline{K} B_x B_\theta}{\bar{r}\sqrt{n}} \cdot \max\{\frac{1}{\sqrt{2}\tau^2}, \frac{1}{\sqrt{\pi}}, \ln\left(\frac{\sqrt{n}}{2\sqrt{2} B_x B_\theta \bar{r}}\right)/\sqrt{2}\}$, then

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq 4\sqrt{2} B_x B_\theta \bar{r} \sqrt{\frac{1}{n} + \frac{8d\ln(1/\delta)}{n^2\varepsilon^2}} = \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right).$$

The requirement on the choice of ε is not restrictive, because it decreases at a rate of $\frac{1}{\sqrt{n}}$.

□

EC.3. Proofs for Section 4

EC.3.1. Lemma EC.7 and Proof

Lemma EC.7. *Let $\mathbb{P}$ be a distribution that satisfies the assumptions in Theorem 8. Denote $\mathsf{z} := \left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^\star\right\|_2$. Then,*

1. *The CDF of demand $F_y(t) = \left(\Phi * f_{\epsilon^\star,1/\mathsf{z}}\right)(t/\mathsf{z})$ is the convolution between CDF Φ of a standard normal random variable and an adjusted PDF $f_{\epsilon^\star,h}(t) := f_{\epsilon^\star}(t/h)/h$.*
2. *The CDF of demand F_y is differentiable and strictly increasing; therefore, $q^* := F_y^{-1}(r) = \arg\min\{t \in \mathbb{R} : F_y(t) \geq r\}$ is unique and well-defined.*

Proof. To show part 1, we need to check the CDF of demand y. Let $f_{\boldsymbol{x}}$ be the PDF of $\boldsymbol{x}$ and $f_{\boldsymbol{z}}$ be the PDF of a standard multivariate Gaussian $\boldsymbol{z} := (z_1, \ldots, z_d) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d\times d})$ with independent coordinates. Rescaling and rotating give $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}$. Denote $\mathsf{z} := \left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^\star\right\|_2$, then we have:

$$\begin{aligned} F_y(t) = \Pr_y[y \leq t] &= \Pr_{\boldsymbol{x},\epsilon}[\langle\boldsymbol{\theta}^\star, \boldsymbol{x}\rangle + \epsilon^\star \leq t] \\ &= \int_{-\infty}^{\infty} \Pr_{\boldsymbol{x}}[\langle\boldsymbol{\theta}^\star, \boldsymbol{x}\rangle \leq t - u] \cdot f_{\epsilon^\star}(u)\, du \\ &= \int_{-\infty}^{\infty} \Pr_{\boldsymbol{z}}\left[\left\langle\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}\right\rangle \leq t - u\right] \cdot f_{\epsilon^\star}(u)\, du && \text{(rescale and rotate)} \\ &= \int_{-\infty}^{\infty} \Pr_{z_1}\left[z_1 \leq (t-u)/\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^\star\right\|_2\right] \cdot f_{\epsilon^\star}(u)\, du && \text{(by rotational invariance of Gaussian } \boldsymbol{z}) \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \Phi\left(\frac{t-u}{\mathsf{z}}\right) \cdot f_{\epsilon^\star}(u)\, du \\
&= \int_{-\infty}^{\infty} \Phi\left(\frac{t}{\mathsf{z}} - v\right) \cdot f_{\epsilon^\star}(v\mathsf{z})\mathsf{z}\, dv && \text{(change variable } v := u/\mathsf{z}\text{)} \\
&= \int_{-\infty}^{\infty} \Phi\left(\frac{t}{\mathsf{z}} - v\right) \cdot f_{\epsilon^\star,1/\mathsf{z}}(v)\, dv && \text{(by definition of } f_{\epsilon^\star,h}\text{)} \\
&= \left(\Phi * f_{\epsilon^\star,1/\mathsf{z}}\right)(t/\mathsf{z}), && \text{(by definition of convolution (3))}
\end{aligned}
$$

where $f_{\epsilon^\star,h}(t) := f_{\epsilon^\star}(t/h)/h$ is an adjusted function of the pdf $f_{\epsilon^\star}$.

As for part 2 of Lemma EC.7, it is a corollary of part 1 by applying the differentiation property of convolution that the derivative of convolution can be obtained by taking derivative first on either function and then doing convolution, i.e., $(f*g)' = (f'*g) = (f*g')$. Therefore, we have

$$
F_y'(t) = \frac{d\left(\Phi * f_{\epsilon^\star,1/\mathsf{z}}\right)(t/\mathsf{z})}{dt} = \left(\Phi' * f_{\epsilon^\star,1/\mathsf{z}}\right)(t/\mathsf{z}) \cdot \frac{1}{\mathsf{z}} > 0, \quad \forall t,
$$

which implies F_y is differentiable and strictly increasing, ensuring the uniqueness of $q^* := F_y^{-1}(r)$.

□

EC.3.2. Proof of Theorem 8

Proof. Before proving the Theorem, we highlight that this proof heavily relies on technical Lemma EC.7 and reasoning therein.

In this proof, we notationally omit the superscript $^{\mathsf{SAA}}$ and the dependence on $\boldsymbol{y}$ in $\widehat{q}^{\mathsf{SAA}}(\boldsymbol{y})$ and simply use $\widehat{q}$ to represent the SAA order quantity. Now, we come to prove Theorem 8. First of all, because of $\bar{r}$-Lipschitz continuity of newsvendor loss function $c(\cdot)$, the excess generalization risk can be upper bounded as

$$
\begin{aligned}
\mathbb{E}_{y,\boldsymbol{y}}\left[c(y-\widehat{q}(\boldsymbol{y}))\right] - \mathbb{E}_{\boldsymbol{x},y}\left[c(y-\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle)\right] &\le \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[|\widehat{q}-\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle|\right]\cdot\bar{r} \\
&\le \mathbb{E}_{\boldsymbol{y}}\left[|\widehat{q}-q^*|\right]\cdot\bar{r} + \mathbb{E}_{\boldsymbol{x}}\left[|q^*-\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle|\right]\cdot\bar{r}, \qquad \text{(EC.46)}
\end{aligned}
$$

where $q^* := F_y^{-1}(r)$ is unique and well-defined by Lemma EC.7 part 2. Since the second term on the r.h.s of (EC.46) is a constant and strictly greater than 0, it remains to control the distance $|\widehat{q}-q^*|$.

By Lemma EC.7 part 2, we know that the CDF F_y is continuous and strictly increasing; thus $F_y^{-1}(F_y(u)) = u$ for any u. Let the bold symbol $\boldsymbol{y}$ denote a dataset $\{y_i\}_{i=1}^n$, then for any $\epsilon > 0$, by standard analysis, we have:

$$
\begin{aligned}
\Pr_{\boldsymbol{y}}\left[\widehat{q}-q^* \ge \epsilon\right] &= \Pr_{\boldsymbol{y}}\left[\widehat{q} \ge F_y^{-1}(F_y(q^*+\epsilon))\right] && \text{(apply } F_y^{-1}(F_y(u)) = u\text{)} \\
&= \Pr_{\boldsymbol{y}}\left[\sum_{i=1}^n \mathbb{1}\left\{y_i \le F_y(q^*+\epsilon)\right\} \le nr\right] \\
&= \Pr\left[\mathsf{Binomial}(n, F_y(q^*+\epsilon)) \le nr\right] \\
&\le \exp\left(-nD_{KL}\left(r \,\|\, F_y(q^*+\epsilon)\right)\right),
\end{aligned}
$$

where the last line follows the classic Chernoff tail bound on Binomial random variable, and $D_{KL}(p_1 \,\|\, p_2) := p_1 \ln\left(\frac{p_1}{p_2}\right) + (1-p_1)\ln\left(\frac{1-p_1}{1-p_2}\right)$ is the Kullback-Leibler (KL) divergence between two Bernoulli random variables with success probabilities p_1 and p_2. Similarly, we can obtain the bound for the other side that

$$\Pr_{\boldsymbol{y}}[\widehat{q} - q^* \le -\epsilon] \le \exp\left(-nD_{KL}\left(r \,\|\, F_y(q^* - \epsilon)\right)\right).$$

Combining both sides results in

$$\Pr_{\boldsymbol{y}}[|\widehat{q} - q^*| \ge \epsilon] \le \exp\left(-nD_{KL}\left(r \,\|\, F_y(q^* + \epsilon)\right)\right) + \exp\left(-nD_{KL}\left(r \,\|\, F_y(q^* - \epsilon)\right)\right). \tag{EC.47}$$

Since F_y is strictly increasing, there must exist $\tau > 0, s > 0$ such that for any q in the punctured τ-neighborhood of q^* defined as $\{q \in \mathbb{R} : |q - q^*| \le \tau, q \ne q^*\}$, the slope $(F_y(q) - F_y(q^*))/(q - q^*) \ge s$. Consequently, if we assume ϵ is small enough so that $\epsilon \in (0, \tau)$, then $q^* + \epsilon$ is in the punctured neighborhood; and

$$F_y(q^* + \epsilon) \ge F_y(q^*) + s\epsilon = r + s\epsilon;$$
$$F_y(q^* - \epsilon) \le F_y(q^*) - s\epsilon = r - s\epsilon.$$

Therefore,

$$\begin{aligned}(\text{EC.47}) &\le \exp\left(-nD_{KL}\left(r \,\|\, r + s\epsilon\right)\right) + \exp\left(-nD_{KL}\left(r \,\|\, r - s\epsilon\right)\right) && \text{(by KL-divergence)}\\ &\le 2\exp\left(-2ns^2\epsilon^2\right),\end{aligned}$$

where the last inequality is by applying Pinsker's inequality that $D_{KL}(a \,\|\, b) \ge 2(a-b)^2$. Setting $2\exp(-2ns^2\epsilon^2) = \gamma$ and solving for ϵ, we get $\epsilon = \frac{1}{s}\sqrt{\frac{\ln(2/\gamma)}{2n}}$. That implies, with probability at least $1-\gamma$ over data sampling process, the order quantity $\widehat{q}$ by SAA satisfies

$$|\widehat{q} - q^*| \le \frac{1}{s}\sqrt{\frac{\ln(2/\gamma)}{2n}},$$

for large enough n. Plugging the above inequality back into (EC.46), we obtain the desired Theorem.

□

EC.3.3. Proof of Proposition 1

Proof. The (ε, δ)-DP is a corollary of Advanced Composition Lemma EC.3, and the values $(\varepsilon_0, \delta_0)$ are set accordingly. By setting $(\varepsilon_0, \delta_0)$ as stated values, the regret for each single product is:

$$\sup_{s \in \mathcal{S}} \mathcal{R}(\mathsf{IndepRun_s}; \mathbb{P}_s) = \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{Sd\ln(S/\delta)}}{n\varepsilon}\right), \tag{EC.48}$$

which directly follows from the regret results of proposed algorithms, see Theorem 2, 4, 6.

To show the total regret $\mathcal{R}(\mathsf{IndepRun};\mathbb{P})$, we take a look at Algorithm OP as an example; results on other algorithms can be obtained from a similar analysis with proper modifications. We first notice that, independently running the S models is equivalent to the following problem,

$$\min_{\boldsymbol{\theta}^{\mathcal{S}}} \frac{1}{n}\sum_{i=1}^{n} \ell_h(\boldsymbol{\theta}^{\mathcal{S}};\boldsymbol{x}_i;\boldsymbol{y}_i) + \lambda \left\|\boldsymbol{\theta}^{\mathcal{S}}\right\|_2 + \frac{\langle \boldsymbol{\theta}^{\mathcal{S}}, \boldsymbol{b}\rangle}{n}, \tag{EC.49}$$

where $\ell_h(\boldsymbol{\theta}^{\mathcal{S}};\boldsymbol{x}_i,\boldsymbol{y}_i) = \sum_{s\in\mathcal{S}} \ell_h(\boldsymbol{\theta}_s;\boldsymbol{x}_i,y_{is})$, which is the loss function associated with $\mathcal{R}(\mathsf{IndepRun};\mathbb{P})$. Noise $\boldsymbol{b}\sim\mathcal{N}(\mathbf{0},\sigma_0^2)$ is a dS-dimensional random vector with $\sigma_0 = L^{\mathcal{S}}\sqrt{8\ln(1/\delta_0)+4\varepsilon_0}/\varepsilon_0$. Thus, upper bounding the total regret amounts to upper bounding the regret of (EC.49). We can show that

1. $\ell_h(\boldsymbol{\theta}^{\mathcal{S}})$ is $\bar{r}^{\mathcal{S}}B_x\sqrt{S} =: L^{\mathcal{S}}$-Lipschitz continuous;
2. $\ell_h(\boldsymbol{\theta}^{\mathcal{S}})$ is $\overline{K}B_x^2\sqrt{S}/h =: \beta^{\mathcal{S}}$-smooth.

This is because, for a pair of estimators $\boldsymbol{\theta}^{\mathcal{S}},\boldsymbol{\theta}^{\mathcal{S}\prime}$ trained on a pair of neighboring datasets, we have

$$\begin{aligned}
\text{(Lipschitzness)}\quad \ell_h(\boldsymbol{\theta}^{\mathcal{S}}) - \ell_h(\boldsymbol{\theta}^{\mathcal{S}\prime}) &= \sum_{s\in\mathcal{S}} [\ell_h(\boldsymbol{\theta}_s) - \ell_h(\boldsymbol{\theta}'_s)] \\
&\le \bar{r}^{\mathcal{S}}B_x \sum_{s\in\mathcal{S}} \|\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s\|_2 && \text{(because } \ell_h \text{ is } \bar{r}^{\mathcal{S}}B_x\text{-Lip cts)} \\
&\le \bar{r}^{\mathcal{S}}B_x\sqrt{S}\left\|\boldsymbol{\theta}^{\mathcal{S}} - \boldsymbol{\theta}^{\mathcal{S}\prime}\right\|_2; && \text{(by Cauchy's inequality)} \\
\text{(Smoothness)}\quad \nabla\ell_h(\boldsymbol{\theta}^{\mathcal{S}}) - \nabla\ell_h(\boldsymbol{\theta}^{\mathcal{S}\prime}) &= \sum_{s\in\mathcal{S}} [\nabla\ell(\boldsymbol{\theta}_s) - \nabla\ell_h(\boldsymbol{\theta}'_s)] \\
&\le \overline{K}B_x^2/h \sum_{s\in\mathcal{S}} \|\boldsymbol{\theta}_s - \boldsymbol{\theta}'_s\|_2 && \text{(because } \ell_h \text{ is } \overline{K}B_x^2/h\text{-smooth)} \\
&\le \overline{K}B_x^2\sqrt{S}/h\left\|\boldsymbol{\theta}^{\mathcal{S}} - \boldsymbol{\theta}^{\mathcal{S}\prime}\right\|_2. && \text{(by Cauchy's inequality)}
\end{aligned}$$

Therefore, following the analysis in the proof of Theorem 2, we can upper bound the total regret as

$$\mathcal{R}(\mathsf{IndepRun};\mathbb{P}) \le \mathcal{O}\left(\underbrace{\frac{(L^{\mathcal{S}})^2}{n\lambda}}_{\text{by uniform stability}} + \underbrace{\frac{dS\sigma_0^2}{n^2\lambda} + \lambda M/B_\theta^2}_{\text{by shrinking ERM}} + \underbrace{\frac{S^3}{\lambda^2 n^2\varepsilon_0^2}}_{\text{approximation error}} \right). \tag{EC.50}$$

The r.h.s of (EC.50) is slightly different from (EC.19). Differences and reasons are summarized below: (i) in the second part, which is by shrinking ERM, the additional S in the numerator of its first term is due to increased dimensionality in $\boldsymbol{b}$; and M/B_θ^2 comes from the upper bound of demand M. (ii) in the third part, which is approximation error, the numerator S^3 originates from two sources: regularizer λ now should satisfy $\lambda \ge \frac{\beta S}{n\varepsilon}$ because of the increase in dimensionality, which modifies the exponent on the l.h.s of (EC.10) to S, thus contributing a S^2; The increase in the smoothness parameter contributes another S.

After getting (EC.50), we can set $\lambda = (\frac{(L^{\mathcal{S}})^2}{n} + \frac{dS\sigma_0^2}{n^2})/(M/B_\theta^2)$. Following the remaining analysis steps for Theorem 2, we get

$$\mathcal{R}(\mathsf{IndepRun};\mathbb{P}) = \mathcal{O}\left(\sqrt{\frac{S}{n}} + \frac{\sqrt{S\cdot Sd\ln(S/\delta)}}{n\varepsilon}\right), \quad \forall\mathbb{P}, \tag{EC.51}$$

if $\varepsilon \ge \Omega(S^{5/4}/n^{1/4})$. Comparing (EC.48) and (EC.51) gives the proposition immediately. □

EC.4. Proofs for Section 5

Lemma EC.8 (Closed-form of $\boldsymbol{\theta}^*$ and $F_{\epsilon|\boldsymbol{x}}(0) = r$ when error term is exogenous). *Suppose demand follows the true linear model $y = \langle \boldsymbol{\theta}^\star, \boldsymbol{x}\rangle + \epsilon^\star(\boldsymbol{x})$, where $\epsilon^\star(\boldsymbol{x}) \equiv \epsilon^\star$ is independent of $\boldsymbol{x}$, and has a CDF $F_{\epsilon^\star}$. We further assume $x_1 \equiv 1$ to incorporate an intercept term. Let $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta})$, and let $F^{-1}_{\epsilon^\star}(r) := \inf\{y \in \mathbb{R} : F_{\epsilon^\star}(y) \geq r\}$. Define a new error term $\epsilon(\boldsymbol{x}) := y - \langle \boldsymbol{\theta}^*, \boldsymbol{x}\rangle$.*

1. *If the CDF $F_{\epsilon^\star}$ is continuous at its r-th quantile, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}^\star + [F^{-1}_{\epsilon^\star}(r), 0, \ldots, 0]^\top$, and $F_{\epsilon|\boldsymbol{x}}(0) = r, \forall \boldsymbol{x}$.*
2. *If the CDF $F_{\epsilon^\star}$ is discontinuous at its r-th quantile, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}^\star + [F^{-1}_{\epsilon^\star}(r), 0, \ldots, 0]^\top$, and $F_{\epsilon|\boldsymbol{x}}(0) \geq r, \forall \boldsymbol{x}$.*

Proof of Lemma EC.8. It is well known that the expected newsvendor cost $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},y}\left[c(y - \boldsymbol{\theta}^\top \boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x},\epsilon^\star}\left[c(\boldsymbol{\theta}^{\star\top}\boldsymbol{x} + \epsilon^\star - \boldsymbol{\theta}^\top \boldsymbol{x})\right]$ is differentiable at any $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\begin{aligned}\nabla \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{x},\epsilon^\star}\left[r \cdot (-\boldsymbol{x}) \cdot \mathbb{1}\left\{\boldsymbol{\theta}^{\star\top}\boldsymbol{x} + \epsilon^\star - \boldsymbol{\theta}^\top \boldsymbol{x} \geq 0\right\} + (1-r)\cdot \boldsymbol{x} \cdot \mathbb{1}\left\{\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^{\star\top}\boldsymbol{x} - \epsilon^\star > 0\right\}\right] \\ &= \mathbb{E}_{\boldsymbol{x},\epsilon^\star}\left[\boldsymbol{x} \cdot \mathbb{1}\left\{\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\theta}^{\star\top}\boldsymbol{x} - \epsilon^\star \geq 0\right\}\right] - r\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x}\right].\end{aligned}$$

1. **Continuous CDF.** Because (i) we assume $x_1 \equiv 1$; and (ii) $F_{\epsilon^\star}$ is continuous at r-th quantile, which implies $F_{\epsilon^\star}(F^{-1}_{\epsilon^\star}(r)) = r$, we can show that the gradient at $\boldsymbol{\theta}^* := \boldsymbol{\theta}^\star + [F^{-1}_{\epsilon^\star}(r), 0, \ldots, 0]$ is $\mathbf{0}$:

$$\begin{aligned}\nabla \mathcal{L}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x} \cdot \mathbb{E}_{\epsilon^\star}\left[\mathbb{1}\left\{F^{-1}_{\epsilon^\star}(r)x_1 - \epsilon^\star \geq 0\right\}\right]\right] - r\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x} \cdot \Pr_{\epsilon^\star}\left[\epsilon^\star \leq F^{-1}_{\epsilon^\star}(r)\right]\right] - r\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] && \text{(by } x_1 \equiv 1\text{)} \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x} \cdot r\right] - r\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] && \text{(by } F_{\epsilon^\star}(F^{-1}_{\epsilon^\star}(r)) = r\text{)} \\ &= \mathbf{0}.\end{aligned}$$

Therefore, $\boldsymbol{\theta}^*$ is an optimal solution to $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$. We next show the argument $F_{\epsilon|\boldsymbol{x}}(0) = r, \forall \boldsymbol{x}$:

$$\begin{aligned}F_{\epsilon|\boldsymbol{x}}(0) = \Pr\left[\epsilon(\boldsymbol{x}) \leq 0\right] &= \Pr\left[\langle \boldsymbol{\theta}^\star, \boldsymbol{x}\rangle + \epsilon^\star(\boldsymbol{x}) - \langle \boldsymbol{\theta}^*, \boldsymbol{x}\rangle \leq 0\right] \\ &= \Pr\left[\epsilon^\star(\boldsymbol{x}) \leq -\langle \boldsymbol{\delta}, \boldsymbol{x}\rangle\right] && \text{(let } \boldsymbol{\delta} := \boldsymbol{\theta}^\star - \boldsymbol{\theta}^*\text{)} \\ &= \Pr\left[\epsilon^\star(\boldsymbol{x}) \leq F^{-1}_{\epsilon^\star}(r)\right] && \text{(since } \boldsymbol{\delta} = [-F^{-1}_{\epsilon^\star}(r), 0, \ldots, 0], \text{ and } x_1 \equiv 1\text{)} \\ &= \Pr\left[\epsilon^\star \leq F^{-1}_{\epsilon^\star}(r)\right] && \text{(by exogeneity } \epsilon^\star(\boldsymbol{x}) \equiv \epsilon^\star\text{)} \\ &= r. && \text{(by continuity of } F_{\epsilon^\star}\text{)}\end{aligned}$$

2. **Discontinuous CDF.** When $F_{\epsilon^\star}$ is not continuous at r-th quantile, following the similar argument on $\nabla \mathcal{L}(\boldsymbol{\theta}^*)$, we know the right derivative of loss function $c(\cdot)$ is $F(F^{-1}_{\epsilon^\star}(r)) - r \geq 0$. On the other hand, its left derivative is smaller than 0. This implies $\boldsymbol{\theta}^*$ is the optimal minimizer. Below shows the quantile of $\epsilon(x)$ at 0:

$$F_{\epsilon|\boldsymbol{x}}(0) = \Pr\left[\epsilon(\boldsymbol{x}) \leq 0\right] = \Pr\left[y - \langle \boldsymbol{\theta}^*, \boldsymbol{x}\rangle\right]$$

$$
\begin{aligned}
&= \Pr\left[\epsilon^\star(\boldsymbol{x}) \le F_{\epsilon^\star}^{-1}(r)\right] \\
&= \Pr\left[\epsilon^\star \le F_{\epsilon^\star}^{-1}(r)\right] \\
&= F_{\epsilon^\star}(F_{\epsilon^\star}^{-1}(r)) \\
&\ge r.
\end{aligned}
$$

The last line can only take inequality because $F_{\epsilon^\star}$ is not continuous at r-th quantile. And by definition of F^{-1}, we know this inequality is true.

□

EC.4.1. Lemma EC.9 and Proof

Lemma EC.9 (Population-Level Minimizers' (Asymptotic) Behavior). *Let* $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$, $\boldsymbol{\theta}_h^* := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_h(\boldsymbol{\theta})$, *and* $\boldsymbol{\theta}_h^{\#*} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_h^{\#}(\boldsymbol{\theta})$, *and* $\boldsymbol{\theta}_h^{\mathsf{OP}*} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_h^{\mathsf{OP}}(\boldsymbol{\theta})$. *Denote* $\boldsymbol{J} := \mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right] \succ \mathbf{0}$. *Then, under Assumption 2, if* $\lambda = o(1)$ *and* $h = o(1)$, *we have*

(i) $\boldsymbol{\theta}_h^* - \boldsymbol{\theta}^* = -\frac{1}{2}h^2\kappa_2\boldsymbol{v} + o(h^2)$, *where* $\boldsymbol{v} = \boldsymbol{J}^{-1}\mathbb{E}_{\boldsymbol{x}}\left[f'_{\epsilon|\boldsymbol{x}}(0)\boldsymbol{x}\right]$;

(ii) $\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^* = -2\lambda\boldsymbol{J}^{-1}\boldsymbol{\theta}_h^* + o(\lambda) + o(h^2)$;

(iii) $\boldsymbol{\theta}_h^{\mathsf{OP}*} = \boldsymbol{\theta}_h^{\#*}$.

Proof. We first express gradient $\nabla\mathcal{L}_h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},y}\left[\left[\mathcal{K}_h\left(\boldsymbol{\theta}^\top\boldsymbol{x} - y\right) - r\right]\cdot\boldsymbol{x}\right]$ and Hessian matrix $\nabla^2\mathcal{L}_h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},y}\left[K_h\left(y - \boldsymbol{\theta}^\top\boldsymbol{x}\right)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]$ for any given $\boldsymbol{\theta}\in\mathbb{R}^d$ as functions of the estimation error $\boldsymbol{\delta} := \boldsymbol{\theta} - \boldsymbol{\theta}^*$. Although these expressions are well-known in the literature (Fernandes et al. 2021), we provide detailed proofs for completeness and for later use. We first note that, conditional on feature $\boldsymbol{x}$, we have

$$
\begin{aligned}
\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}\left(\frac{\boldsymbol{\delta}^\top\boldsymbol{x} - \epsilon}{h}\right)\Big|\boldsymbol{x}\right] &= \int_{-\infty}^{\infty}\mathcal{K}\left(\frac{\boldsymbol{\delta}^\top\boldsymbol{x} - t}{h}\right)dF_{\epsilon|\boldsymbol{x}}(t) \\
&= \frac{1}{h}\int_{-\infty}^{\infty}F_{\epsilon|\boldsymbol{x}}(t)K\left(\frac{\boldsymbol{\delta}^\top\boldsymbol{x} - t}{h}\right)dt && \text{(integration by parts)} \\
&= \int_{-\infty}^{\infty}F_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}^\top\boldsymbol{x} - uh)K(u)\,du && (\text{let } u = (\boldsymbol{\delta}^\top\boldsymbol{x} - t)/h) \\
&= r + \int_{-\infty}^{\infty}\left[F_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}^\top\boldsymbol{x} - uh) - F_{\epsilon|\boldsymbol{x}}(0)\right]K(u)\,du, \quad (F_{\epsilon|\boldsymbol{x}}(0) = r \text{ and } \int K = 1)
\end{aligned}
\tag{EC.52}
$$

and

$$
\begin{aligned}
\mathbb{E}_{y|\boldsymbol{x}}\left[K_h(y - \boldsymbol{\theta}^\top\boldsymbol{x})\big|\boldsymbol{x}\right] &= \int_{-\infty}^{\infty}\frac{1}{h}K\left(\frac{\boldsymbol{\theta}^{*\top}\boldsymbol{x} + t - \boldsymbol{\theta}^\top\boldsymbol{x}}{h}\right)dF_{\epsilon|\boldsymbol{x}}(t) \\
&= \int_{-\infty}^{\infty}K(u)f_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}^\top\boldsymbol{x} + uh)\,du\,. && \left(\text{let } u = \frac{t - \boldsymbol{\delta}^\top\boldsymbol{x}}{h}\right)
\end{aligned}
\tag{EC.53}
$$

Following (EC.52) and (EC.53), the gradient and Hessian matrix can be expressed as

$$
\nabla\mathcal{L}_h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}_h\left(\boldsymbol{\delta}^\top\boldsymbol{x} - \epsilon\right) - r\right]\cdot\boldsymbol{x}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty}K(u)\cdot\int_0^{\boldsymbol{\delta}^\top\boldsymbol{x} - uh}f_{\epsilon|\boldsymbol{x}}(t)\,dt\,du\cdot\boldsymbol{x}\right]; \tag{EC.54}
$$

$$\nabla^2 \mathcal{L}_h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[K_h(\epsilon - \boldsymbol{\delta}^\top \boldsymbol{x})\right] \cdot \boldsymbol{x}\boldsymbol{x}^\top\right] = \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \cdot f_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}^\top \boldsymbol{x} + uh)\, du \cdot \boldsymbol{x}\boldsymbol{x}^\top\right]. \tag{EC.55}$$

With above expressions, we are ready to prove the three statements in the Lemma.

(i) According to (EC.54), the gradient at $\boldsymbol{\theta}^*$ (which implies $\boldsymbol{\delta} = \mathbf{0}$) is

$$\begin{aligned}
\nabla \mathcal{L}_h(\boldsymbol{\theta}^*) &= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \cdot \int_0^{-uh} f_{\epsilon|\boldsymbol{x}}(t) - f_{\epsilon|\boldsymbol{x}}(0)\, dt\, du \cdot \boldsymbol{x}\right] \quad \left(\int K(u) \cdot (-uh) f_{\epsilon|\boldsymbol{x}}(0)\, du = 0\right) \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \cdot \int_0^{-uh} t f'_{\epsilon|\boldsymbol{x}}(0) + o(t)\, dt\, du \cdot \boldsymbol{x}\right] \qquad \text{(by Taylor expansion)} \\
&= \frac{1}{2} h^2 \kappa_2 \mathbb{E}_{\boldsymbol{x}}\left[f'_{\epsilon|\boldsymbol{x}}(0)\boldsymbol{x}\right] + o(h^2).
\end{aligned}$$

And according to (EC.55), the Hessian matrix at $\boldsymbol{\theta}^*$ is

$$\begin{aligned}
\nabla^2 \mathcal{L}_h(\boldsymbol{\theta}^*) &= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) f_{\epsilon|\boldsymbol{x}}(uh)\, du \cdot \boldsymbol{x}\boldsymbol{x}^\top\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \cdot \left[f_{\epsilon|\boldsymbol{x}}(0) + f'_{\epsilon|\boldsymbol{x}}(0)uh + o(uh)\right] du \cdot \boldsymbol{x}\boldsymbol{x}^\top\right] \qquad \text{(by Taylor expansion)} \\
&= \boldsymbol{J} + o(h),
\end{aligned}$$

where $\boldsymbol{J} := \mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right]$. The Taylor expansions can be applied because of our assumption that $f_{\epsilon|\boldsymbol{x}}$ is differentiable at 0. We further notice that, with $\boldsymbol{\theta}_h^*$, the population-level gradient is $\nabla \mathcal{L}_h(\boldsymbol{\theta}_h^*) = \mathbf{0}$. Let $\boldsymbol{\delta}_h^* := \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^*$, then

$$-\nabla \mathcal{L}_h(\boldsymbol{\theta}^*) = \nabla \mathcal{L}_h(\boldsymbol{\theta}_h^*) - \nabla \mathcal{L}_h(\boldsymbol{\theta}^*) = \nabla^2 \mathcal{L}_h(\boldsymbol{\theta}^*)\boldsymbol{\delta}_h^* + o(\|\boldsymbol{\delta}_h^*\|_2),$$

where the first equality is due to First-Order-Condition that $\nabla \mathcal{L}_h(\boldsymbol{\theta}_h^*) = \mathbf{0}$, and the second equality follows from Taylor expansion of $\nabla \mathcal{L}_h$ at $\boldsymbol{\theta}^*$. By omitting the higher-order infinitesimal $o(\|\boldsymbol{\delta}_h^*\|_2)$, we have $\boldsymbol{\delta}_h^* = -\left(\nabla^2 \mathcal{L}_h(\boldsymbol{\theta}^*)\right)^{-1} \nabla \mathcal{L}_h(\boldsymbol{\theta}^*)$. Replacing the gradient and Hessian matrix with their expressions derived above gives the first statement.

(ii) As for the second part, from First-Order-Conditions, we know $\nabla \mathcal{L}_h(\boldsymbol{\theta}_h^*) = \mathbf{0}$ and $\nabla \mathcal{L}_h^{\#*}(\boldsymbol{\theta}_h^{\#*}) = \mathbf{0}$. According to (EC.54), we can express

$$\begin{aligned}
\nabla \mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\#*}) &= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \int_0^{(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}^*)^\top \boldsymbol{x} - uh} f_{\epsilon|\boldsymbol{x}}(t)\, dt\, du \cdot \boldsymbol{x}\right] + 2\lambda \boldsymbol{\theta}_h^{\#*} \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \int_0^{\boldsymbol{\delta}_h^{*\top} \boldsymbol{x} - uh} f_{\epsilon|\boldsymbol{x}}(t)\, dt\, du \cdot \boldsymbol{x}\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \int_{\boldsymbol{\delta}_h^{*\top} \boldsymbol{x} - uh}^{\boldsymbol{\delta}_h^{*\top} \boldsymbol{x} - uh + (\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top \boldsymbol{x}} f_{\epsilon|\boldsymbol{x}}(t)\, dt\, du \cdot \boldsymbol{x}\right] + 2\lambda \boldsymbol{\theta}_h^{\#*} \\
&= \nabla \mathcal{L}_h(\boldsymbol{\theta}_h^*) + \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u) \int_{\boldsymbol{\delta}_h^{*\top} \boldsymbol{x} - uh}^{\boldsymbol{\delta}_h^{*\top} \boldsymbol{x} - uh + (\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top \boldsymbol{x}} f_{\epsilon|\boldsymbol{x}}(t)\, dt\, du \cdot \boldsymbol{x}\right] + 2\lambda \boldsymbol{\theta}_h^{\#*}.
\end{aligned} \tag{EC.56}$$

Since $\|\boldsymbol{\delta}_h^*\|_2 = O(h^2)$ and $\boldsymbol{\theta}_h^{\#*} \nearrow \boldsymbol{\theta}_h^*$ as $\lambda \searrow 0$, the upper limit $u(\boldsymbol{\delta}_h^*) := \boldsymbol{\delta}_h^{*\top}\boldsymbol{x} - uh + (\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top \boldsymbol{x}$ and lower limit $l(\boldsymbol{\delta}_h^*) := \boldsymbol{\delta}_h^{*\top}\boldsymbol{x} - uh$ of the most inner integration on the r.h.s. of (EC.56) will tend to 0; thus we can approximate the integration by Taylor expansion at 0:

$$\begin{aligned}
\int_{l(\boldsymbol{\delta}_h^*)}^{u(\boldsymbol{\delta}_h^*)} f_{\epsilon|\boldsymbol{x}}(t)\,dt &= \left[F_{\epsilon|\boldsymbol{x}}(0) + f_{\epsilon|\boldsymbol{x}}(0)u(\boldsymbol{\delta}_h^*) + \frac{1}{2}f'_{\epsilon|\boldsymbol{x}}(0)u^2(\boldsymbol{\delta}_h^*) + o(u^2(\boldsymbol{\delta}_h^*))\right] \\
&\quad - \left[F_{\epsilon|\boldsymbol{x}}(0) + f_{\epsilon|\boldsymbol{x}}(0)l(\boldsymbol{\delta}_h^*) + \frac{1}{2}f'_{\epsilon|\boldsymbol{x}}(0)l^2(\boldsymbol{\delta}_h^*) + o(l^2(\boldsymbol{\delta}_h^*))\right] \\
&= f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x} + \frac{1}{2}f'_{\epsilon|\boldsymbol{x}}(0)\left[2(\boldsymbol{\delta}_h^{*\top}\boldsymbol{x} - uh)(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x} + \left[(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x}\right]^2\right] \\
&\quad + o(h^2) + o(\left\|\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*\right\|_2^2) \\
&= f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x} - uhf'_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x} + \frac{1}{2}f'_{\epsilon|\boldsymbol{x}}(0)\left[(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x}\right]^2 \\
&\quad + o(h^2) + o(\left\|\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*\right\|_2^2),
\end{aligned}$$

where the last equality is due to $\boldsymbol{\delta}_h^{*\top}\boldsymbol{x}(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x} = O(h^2)o(1) = o(h^2)$. Plugging above value into (EC.56), we can obtain that

$$\begin{aligned}
\nabla\mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\#*}) &= \mathbb{E}_{\boldsymbol{x}}\left[\left[f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x} + \frac{1}{2}f'_{\epsilon|\boldsymbol{x}}(0)\left[(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x}\right]^2\right]\cdot\boldsymbol{x}\right] + 2\lambda\boldsymbol{\theta}_h^{\#*} \\
&\quad + o(h^2) + o(\left\|(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)\right\|_2^2) \\
&= \boldsymbol{J}\cdot(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*) + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[f'_{\epsilon|\boldsymbol{x}}(0)\left[(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*)^\top\boldsymbol{x}\right]^2\cdot\boldsymbol{x}\right] + 2\lambda\boldsymbol{\theta}_h^{\#*} + o(h^2) + o(\left\|\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*\right\|_2^2) \\
&= \boldsymbol{J}\cdot(\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*) + 2\lambda\boldsymbol{\theta}_h^{\#*} + o(h^2) + o(\left\|\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*\right\|_2) = \mathbf{0}.
\end{aligned}$$

Rearranging the above expression gives $\boldsymbol{\theta}_h^{\#*} = (\boldsymbol{I} + 2\lambda\boldsymbol{J}^{-1})^{-1}\boldsymbol{\theta}_h^* + o(h^2) + o(\|\boldsymbol{\theta}_h^{\#*} - \boldsymbol{\theta}_h^*\|_2)$. Then the second part of this Lemma directly follows from inverse matrix approximation with second-order Neumann series that $(\boldsymbol{I} + A)^{-1} \cong I - A + A^2$ if $A \to \mathbf{0}$.

(iii) The proof for the third part of this Lemma is straightforward and is shown below:

$$\begin{aligned}
\boldsymbol{\theta}_h^{\mathsf{OP}*} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathcal{L}_h^{\mathsf{OP}}(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathbb{E}_{(\boldsymbol{x},y),\boldsymbol{b}}\left[\ell_h(\boldsymbol{\theta};\boldsymbol{x},y) + \lambda\|\boldsymbol{\theta}\|_2^2 + \frac{\boldsymbol{b}^\top\boldsymbol{\theta}}{n}\right] \\
&= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathbb{E}_{(\boldsymbol{x},y)}\left[\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\right] + \lambda\|\boldsymbol{\theta}\|_2^2 \\
&= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathcal{L}_h^{\#}(\boldsymbol{\theta}) = \boldsymbol{\theta}_h^{\#*}.
\end{aligned}$$

□

Lemma EC.10 (Lemma 9.21 in Wainwright 2019, rephrased). *Denote* $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}}\widehat{\mathcal{L}}(\boldsymbol{\theta})$. *For a given radius* r_0, *if* $\widehat{\mathcal{L}}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*) > 0$, $\forall\boldsymbol{\delta} \in \{\boldsymbol{\delta}\in\mathbb{R}^d : \|\boldsymbol{\delta}\|_2 = r_0\}$, *then* $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \le r_0$.

EC.4.2. Proof of Proposition 2

Proof. Recall that, for any upcoming feature vector $\boldsymbol{x}$, the private order quantity is $\widehat{q}_h^{\mathsf{OP}} = \left\langle \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}, \boldsymbol{x} \right\rangle$; thus, we only need to show $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} = \boldsymbol{\theta}^* - 2\lambda\boldsymbol{u} - \frac{1}{2}h^2\kappa_2\boldsymbol{v} + \mathcal{O}_p\left(\frac{1}{\sqrt{n}} + \frac{\sigma}{n}\right)$, with $\boldsymbol{u}, \boldsymbol{v}$, and $\boldsymbol{J}$ as stated. In the following analysis, we will ignore the infinitesimals of λ and h^2 for conciseness. Let symbol $\mathbb{P}$ represents dataset sampling randomness, symbol $\mathbb{Q}$ represents the algorithm's randomness, and let $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) := \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \mathcal{L}_h^{\#}(\boldsymbol{\theta}) + \boldsymbol{b}^\top\boldsymbol{\theta}/n$ for any given $\boldsymbol{b}$. From Lemma EC.9, we know that, by ignoring $o(\lambda)$ and $o(h^2)$, $\boldsymbol{\theta}_h^{\mathsf{OP}*} - \boldsymbol{\theta}^* = -2\lambda\boldsymbol{u} - \frac{1}{2}h^2\kappa_2\boldsymbol{v}$. Therefore, it remains to show $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} \xrightarrow{\mathbb{P}\times\mathbb{Q}} \boldsymbol{\theta}_h^{\mathsf{OP}*}$ at the given rate $\frac{1}{\sqrt{n}} + \frac{\sigma}{n}$, or equivalently to show (i) $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b})$ at rate $\frac{1}{\sqrt{n}}$ for any given $\boldsymbol{b}$; and (ii) $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) \xrightarrow{\mathbb{Q}} \boldsymbol{\theta}_h^{\mathsf{OP}*}$ at rate $\frac{\sigma}{n}$ simultaneously.

We first fix a noise vector $\boldsymbol{b}$ and show the former convergence $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b})$. Let us treat $\ell_h(\boldsymbol{\theta}) + \boldsymbol{b}^\top\boldsymbol{\theta}/n$ as a new loss function of interest. Because we assume that feature space $\mathcal{X}$ and parameter space $\mathcal{C}$ are compact, and assume that the new loss function has a finite expectation. Therefore, by uniform laws of large numbers, we have

$$\sup_{\boldsymbol{\theta}\in\mathcal{C}} \left| \left(\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta}) + \boldsymbol{b}^\top\boldsymbol{\theta}/n\right) - \left(\mathcal{L}_h^{\#}(\boldsymbol{\theta}) + \boldsymbol{b}^\top\boldsymbol{\theta}/n\right) \right| \xrightarrow{\mathbb{P}} 0.$$

Furthermore, we know that function $\mathcal{L}_h^{\#}(\boldsymbol{\theta}) + \boldsymbol{b}^\top\boldsymbol{\theta}/n$ is strongly convex, implying a unique and well-separated minimizer $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b})$ for any given noise $\boldsymbol{b}$. Consequently, the sequence of empirical minimizers $\{\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\}_{n=1}^{\infty}$ will converge to $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b})$, i.e., $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b})$. More importantly, the convergence rate can be shown to be $\mathcal{O}_p(\frac{1}{\sqrt{n}})$ by standard Bayes's risk analysis, which is tight.

Now, we come to show the latter convergence $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) \xrightarrow{\mathbb{Q}} \boldsymbol{\theta}_h^{\mathsf{OP}*}$. We conjecture that $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) = \boldsymbol{\theta}_h^{\mathsf{OP}*} + \mathcal{O}_p(\frac{\sigma}{n})$, and highlight that the stochastic convergence rate $\mathcal{O}_p\left(\frac{\sigma}{n}\right)$ is *NOT* tight; instead, infinitesimals of λ and h^2 should be included to get a tighter bound $\mathcal{O}_p\left(\frac{\sigma}{n} + o(\lambda) + o(h^2)\right)$. However, we deliberately ignore them for conciseness. To show $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) = \boldsymbol{\theta}_h^{\mathsf{OP}*} + \mathcal{O}_p(\frac{\sigma}{n})$, by the definition of stochastic convergence, it is equivalent to show

$$\Pr_{\boldsymbol{b}}\left[\left\|\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) - \boldsymbol{\theta}_h^{\mathsf{OP}*}\right\|_2 \geq \epsilon\frac{\sigma}{n}\right] \leq C_\epsilon, \quad \forall\epsilon > 0, \tag{EC.57}$$

with a constant C_ϵ that depends on ϵ. Therefore, we only need to find a proper C_ϵ. By a contrapositive statement of Lemma 9.21 in Wainwright (2019) (Lemma EC.10), i.e., if $\left\|\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) - \boldsymbol{\theta}_h^{\mathsf{OP}*}\right\|_2 \geq \epsilon\frac{\sigma}{n}$, then

$$\inf_{\boldsymbol{\delta}\in\partial\mathcal{B}(\epsilon\sigma/n)} \left(\mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\mathsf{OP}*} + \boldsymbol{\delta}) + \frac{\boldsymbol{b}^\top(\boldsymbol{\theta}_h^{\mathsf{OP}*} + \boldsymbol{\delta})}{n}\right) - \left(\mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\mathsf{OP}*}) + \frac{\boldsymbol{b}^\top\boldsymbol{\theta}_h^{\mathsf{OP}*}}{n}\right) \leq 0, \tag{EC.58}$$

it suffices to upper bound the probability on the l.h.s. of (EC.57) by the occurrence probability of (EC.58). Working on the l.h.s. of (EC.58), we have

$$\begin{aligned}
\text{l.h.s. of (EC.58)} &= \inf_{\boldsymbol{\delta}\in\partial\mathcal{B}(\epsilon\sigma/n)} \mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\mathsf{OP}*} + \boldsymbol{\delta}) - \mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\mathsf{OP}*}) + \frac{\boldsymbol{b}^\top\boldsymbol{\delta}}{n} \\
&= \inf_{\boldsymbol{\delta}\in\partial\mathcal{B}(\epsilon\sigma/n)} \int_0^1 \nabla^\top\mathcal{L}_h^{\#}(\boldsymbol{\theta}_h^{\mathsf{OP}*} + t\boldsymbol{\delta})\,dt\cdot\boldsymbol{\delta} + \frac{\boldsymbol{b}^\top\boldsymbol{\delta}}{n},
\end{aligned}$$

$$= \inf_{\boldsymbol{\delta}\in\partial\mathcal{B}(\epsilon\sigma/n)} \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{J}\boldsymbol{\delta} + \lambda\|\boldsymbol{\delta}\|_2^2 + \frac{\boldsymbol{b}^\top\boldsymbol{\delta}}{n} \qquad \text{(by (EC.54), Lemma EC.9, and Taylor expansion)}$$

$$> \lambda_{\min}\cdot\frac{\epsilon^2\sigma^2}{2n^2} - \frac{\|\boldsymbol{b}\|_2\,\epsilon\sigma}{n^2}. \qquad (\lambda_{\min}>0 \text{ is the minimal eigenvalue of } \boldsymbol{J})$$

The above analysis implies

$$\text{l.h.s. of (EC.57)} < \Pr_{\boldsymbol{b}}\left[\lambda_{\min}\cdot\frac{\epsilon^2\sigma^2}{2n^2} - \frac{\|\boldsymbol{b}\|_2\,\epsilon\sigma}{n^2}\le 0\right] = \Pr_{\boldsymbol{b}}\left[\frac{\|\boldsymbol{b}\|_2}{\sigma}\ge\lambda_{\min}\frac{\epsilon}{2}\right]$$
$$\le 2\exp\left(-\lambda_{\min}^2\epsilon^2/(8d)\right) =: C_\epsilon, \quad \forall\epsilon>0,$$

where the last inequality is from Gaussian vector's tail bound since $\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})$. The constant C_ϵ defined above validates (EC.57), further confirming our conjecture $\boldsymbol{\theta}_h^{\mathsf{OP}*}(\boldsymbol{b}) = \boldsymbol{\theta}_h^{\mathsf{OP}*} + \mathcal{O}_p(\frac{\sigma}{n})$.

Combining above analyses together, we get $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} = \boldsymbol{\theta}^* - 2\lambda\boldsymbol{u} - \frac{1}{2}h^2\kappa_2\boldsymbol{v} + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}+\frac{\sigma}{n}\right)$, which immediately results in the desired proposition.

□

EC.4.3. Lemma EC.11 and Proof

Lemma EC.11 (Asymptotic Covariance Matrix of $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$). *Suppose Assumption 2 holds and $\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle\ge 0$ for any $\boldsymbol{x}$. If we set $\lambda = o(1)$ and $h = o(1)$, then the variance of estimator obtained from Algorithm* OP *satisfies*

$$\mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\sqrt{n}\cdot\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right] \succcurlyeq \boldsymbol{J}^{-1}\boldsymbol{\Sigma}_r\boldsymbol{J}^{-1} - c_1\lambda\boldsymbol{J}^{-1}\boldsymbol{\Sigma}_r\boldsymbol{J}^{-1} - 2h\mathsf{K}\boldsymbol{J}^{-1} + \frac{\sigma^2}{n}\boldsymbol{J}^{-1}\boldsymbol{J}^{-1},$$

where $\boldsymbol{J} := \mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0)\cdot\boldsymbol{x}\boldsymbol{x}^T\right]\succ\boldsymbol{0}$, $\boldsymbol{\Sigma}_r := r(1-r)\boldsymbol{\Sigma}\succ\boldsymbol{0}$, $\mathsf{K} := \int_{-\infty}^{\infty} uK(u)\cdot\int_{-\infty}^{u}K(v)\,dv\cdot du\in(0,\kappa_1)$, and $c_1>0$ is a constant. Higher-order infinitesimals of h and λ are omitted. The randomness is taken over both dataset $\mathcal{D}$ generating process and algorithm's randomness.

Proof of Lemma EC.11. Because the following four facts hold for any $h,\lambda>0$:

(i) the loss function $\ell_h^{\mathsf{OP}}(\boldsymbol{\theta};\boldsymbol{x},y) := \ell_h(\boldsymbol{\theta};\boldsymbol{x},y) + \lambda\|\boldsymbol{\theta}\|_2^2 + \frac{\boldsymbol{b}^\top\boldsymbol{\theta}}{n}$ is differentiable in $\boldsymbol{\theta}$ and has a finite expected gradient $\mathbb{E}\left[\left\|\nabla\ell_h^{\mathsf{OP}}(\boldsymbol{\theta}_h^{\mathsf{OP}*})\right\|_2^2\right]<\infty$ at optimal $\boldsymbol{\theta}_h^{\mathsf{OP}*}$ if $\sigma\le\mathcal{O}(n)$. Furthermore, the function ℓ_h^{OP} is also $(\frac{\overline{K}B_x^2}{h}+2\lambda)$-smooth, implying that

$$\left\|\nabla\ell_h^{\mathsf{OP}}(\boldsymbol{\theta}_1) - \nabla\ell_h^{\mathsf{OP}}(\boldsymbol{\theta}_2)\right\|_2 \le (\frac{\overline{K}B_x^2}{h}+2\lambda)\cdot\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2, \quad \forall\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;$$

(ii) the population-level function $\mathcal{L}_h^{\mathsf{OP}}(\boldsymbol{\theta})$'s Hessian matrix $\boldsymbol{H}_h^{\#}(\boldsymbol{\theta}) := \nabla^2\mathcal{L}_h(\boldsymbol{\theta}) + 2\lambda\boldsymbol{I}$ is positive definite at optimal $\boldsymbol{\theta}_h^{\mathsf{OP}*}$ (or equivalently at $\boldsymbol{\theta}_h^{\#*}$, see Lemma EC.9), i.e., $\boldsymbol{H}_h^{\#} := \boldsymbol{H}_h^{\#}(\boldsymbol{\theta}_h^{\#*}) = \mathbb{E}_{\boldsymbol{x},y}\left[K_h(y-\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x})\cdot\boldsymbol{x}\boldsymbol{x}^\top\right] + 2\lambda\boldsymbol{I}\succ\boldsymbol{0}$;

(iii) the empirical gradient satisfies $\nabla\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) = \boldsymbol{0}$;

(iv) the private estimator $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} \xrightarrow{\mathbb{P}^n\times\mathbb{Q}} \boldsymbol{\theta}_h^{\mathsf{OP}*} = \boldsymbol{\theta}_h^{\#*}$,

by Van der Vaart (2000, Theorem 5.23), the private estimator $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$ admits a Bahadur representation:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}} - \boldsymbol{\theta}_h^{\mathsf{OP}*}) = -\boldsymbol{H}_h^{\#-1}\cdot\sqrt{n}\nabla\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta}_h^{\mathsf{OP}*}) + o_p(1), \qquad \text{(EC.59)}$$

where the stochastic convergence $o_p(1)$ is with respect to both data sampling and algorithm's internal randomness. To obtain (EC.59), we need to check the validity of preceding four facts. The first three facts are easy to check. The fourth fact is a corollary of Lemma EC.9 and is an intermediate result in the proof of Proposition 2.

Hence, by (EC.59), the asymptotic covariance matrix of $\sqrt{n}\,\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$ is

$$\mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\sqrt{n}\,\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right] = \boldsymbol{H}_h^{\#-1}\cdot\mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\sqrt{n}\nabla\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta}_h^{\mathsf{OP}*})\right]\cdot\boldsymbol{H}_h^{\#-1}. \tag{EC.60}$$

Since the middle term of the preceding sandwich expression is

$$\begin{aligned}\mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\sqrt{n}\nabla\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta}_h^{\mathsf{OP}*})\right] &= \mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\sqrt{n}\nabla\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta}_h^{\#*})\right]\\ &= n\cdot\mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\frac{1}{n}\sum_{i=1}^{n}\nabla\ell_h(\boldsymbol{\theta}_h^{\#*};\boldsymbol{x}_i,y_i)+2\lambda\boldsymbol{\theta}_h^{\#*}+\frac{\boldsymbol{b}}{n}\right]\\ &= \mathsf{Var}_{(\boldsymbol{x},y)}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right]+\frac{\sigma^2}{n}\boldsymbol{I},\end{aligned}$$

it remains to figure out $\boldsymbol{H}_h^{\#-1}$ and $\mathsf{Var}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right]$. We will show later that

$$\boldsymbol{H}_h^{\#-1} = \boldsymbol{J}^{-1}-2\lambda\cdot\boldsymbol{J}^{-1}\left(\boldsymbol{I}-\boldsymbol{V}'\right)\boldsymbol{J}^{-1}+o(\lambda)+o(h)\succ\mathbf{0}; \tag{EC.61}$$

$$\mathsf{Var}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right] = \boldsymbol{\Sigma}_r+\lambda\cdot(4r-2)\boldsymbol{V}-h\cdot 2\mathsf{K}\boldsymbol{J}+o(\lambda)+o(h)\succ\mathbf{0}, \tag{EC.62}$$

where matrices $\boldsymbol{J}:=\mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]\succ\mathbf{0}$, $\boldsymbol{V}:=\mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]\succ\mathbf{0}$, $\boldsymbol{V}':=\mathbb{E}_{\boldsymbol{x}}\left[f'_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]$, and $\boldsymbol{\Sigma}_r:=r(1-r)\boldsymbol{\Sigma}\succ\mathbf{0}$; constant $\mathsf{K}:=\int_{-\infty}^{\infty}uK(u)\cdot\int_{-\infty}^{u}K(v)\,dv\cdot du\in(0,\kappa_1)$ is derived from kernel function K and independent of h by definition. Substituting preceding two expressions into (EC.60), we obtain by dropping dominated terms,

$$\begin{aligned}\mathsf{Var}_{\mathcal{D},\boldsymbol{b}}\left[\sqrt{n}\,\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right] &= \boldsymbol{J}^{-1}\boldsymbol{\Sigma}_r\boldsymbol{J}^{-1}-\lambda\cdot\boldsymbol{J}^{-1}\left[(4r-2)\boldsymbol{V}+4(\boldsymbol{I}-\boldsymbol{V}')\boldsymbol{J}^{-1}\boldsymbol{\Sigma}_r\right]\boldsymbol{J}^{-1}\\ &\quad -h\cdot 2\mathsf{K}\boldsymbol{J}^{-1}+\frac{\sigma^2}{n}\cdot\boldsymbol{J}^{-1}\boldsymbol{J}^{-1}+o(\lambda)+o(h)\\ &\succcurlyeq \boldsymbol{J}^{-1}\boldsymbol{\Sigma}_r\boldsymbol{J}^{-1}-\lambda\cdot c_1\boldsymbol{J}^{-1}\boldsymbol{\Sigma}_r\boldsymbol{J}^{-1}-h\cdot 2\mathsf{K}\boldsymbol{J}^{-1}+\frac{\sigma^2}{n}\boldsymbol{J}^{-1}\boldsymbol{J}^{-1}+o(\lambda)+o(h),\quad\text{as }\lambda,h\to 0,\end{aligned}$$

for some positive constant c_1. Thus, we get the desired statement.

To complete the proof, we only need to derive (EC.61) and (EC.62). Denote $\boldsymbol{\delta}_h^{\#*}:=\boldsymbol{\theta}_h^{\#*}-\boldsymbol{\theta}^*$.

- We first derive (EC.62). By noticing that

$$\mathsf{Var}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right] = \mathbb{E}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\cdot\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})^\top\right]-\mathbb{E}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right]\cdot\mathbb{E}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right]^\top, \tag{EC.63}$$

  we can separately address the two terms. Following (EC.54), we have

$$\mathbb{E}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right] = \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty}K(u)\cdot\int_{0}^{\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}-uh}f_{\epsilon|\boldsymbol{x}}(t)\,dt\,du\cdot\boldsymbol{x}\right]\qquad(\boldsymbol{\delta}_h^{\#*}:=\boldsymbol{\theta}_h^{\#*}-\boldsymbol{\theta}^*)$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u)\cdot\int_{0}^{-2\lambda(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}-uh+o(\lambda)+o(h^2)} f_{\epsilon|\boldsymbol{x}}(t)\,dt\,du\cdot\boldsymbol{x}\right] \qquad \text{(by Lemma EC.9)}$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u)\cdot\left[f_{\epsilon|\boldsymbol{x}}(0)(-2\lambda(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x})+o(\lambda)+o(h^2)\right]du\cdot\boldsymbol{x}\right] \quad \text{(Taylor expansion)} \tag{EC.64}$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[-2\lambda f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}\cdot\boldsymbol{x}\right]+o(\lambda)+o(h^2)$$

$$= -2\lambda\boldsymbol{\theta}^*+o(\lambda)+o(h^2). \tag{EC.65}$$

We then come to address $\mathbb{E}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\cdot\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})^\top\right]$, which is equal to

$$\mathbb{E}\left[(\mathcal{K}_h(\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x}-y)-r)^2\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]=\mathbb{E}_{\boldsymbol{x}}\left[\left(\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}_h^2(\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x}-y)-2r\mathcal{K}_h(\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x}-y)\right]+r^2\right)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]. \tag{EC.66}$$

We focus on the inner expectation term $\mathbb{E}_{\epsilon|\boldsymbol{x}}[\cdot]$ conditional on $\boldsymbol{x}$ on the r.h.s. of (EC.66). Following a similar idea for (EC.64), we can show that,

$$\begin{aligned}
\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}_h^2(\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x}-y)\right] &= \int_{-\infty}^{\infty} F_{\epsilon|\boldsymbol{x}}(t)\cdot 2\mathcal{K}_h(\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}-t)K_h(\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}-t)\,dt \quad \text{(integration by parts)}\\
&= \int_{-\infty}^{\infty} F_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}-uh)\cdot 2\mathcal{K}_h(uh)K(u)\,du\\
&= \int_{-\infty}^{\infty}\left[\int_0^{\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}-uh} f_{\epsilon|\boldsymbol{x}}(t)\,dt\right]\cdot 2\mathcal{K}_h(uh)K(u)\,du + r\underbrace{\int_{-\infty}^{\infty} 2\mathcal{K}_h(uh)K(u)\,du}_{=1}\\
&= r+\int_{-\infty}^{\infty}\left(-2\lambda f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}-f_{\epsilon|\boldsymbol{x}}(0)uh+o(\lambda)+o(h^2)\right)\cdot 2\mathcal{K}_h(uh)K(u)\,du\\
&= r-2\lambda f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}-2hf_{\epsilon|\boldsymbol{x}}(0)\mathsf{K}+o(\lambda)+o(h^2),
\end{aligned} \tag{EC.67}$$

where $\mathsf{K}=\int_{-\infty}^{\infty}u\mathcal{K}_h(uh)K(u)\,du\in(0,\kappa_1)$ is a real value derived from kernel function K and independent of h by definition. Additionally, following the analysis on $\mathbb{E}\left[\nabla\ell_h(\boldsymbol{\theta}_h^{\#*})\right]$ and (EC.64), we have

$$\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}_h(\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x}-y)\right]=r-2\lambda f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}+o(\lambda)+o(h^2). \tag{EC.68}$$

Plugging (EC.67) and (EC.68) back into (EC.66) gives

$$(\text{EC.66})=\boldsymbol{\Sigma}_r+2\lambda(2r-1)\boldsymbol{V}-2h\mathsf{K}\boldsymbol{J}+o(\lambda)+o(h)\succcurlyeq\mathbf{0} \tag{EC.69}$$

where $\boldsymbol{\Sigma}_r:=r(1-r)\boldsymbol{\Sigma}\succ\mathbf{0}$, and $\boldsymbol{V}:=\mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]\succ\mathbf{0}$ as $\boldsymbol{\theta}^{*\top}\boldsymbol{x}$ is always positive by assumption. Lastly, inserting (EC.65) and (EC.69) into (EC.63) gives (EC.62).

- We next analyze $\boldsymbol{H}_h^{\#-1}$:

$$\begin{aligned}
\boldsymbol{H}_h^{\#}-2\lambda\boldsymbol{I} &= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[K_h(y-\boldsymbol{\theta}_h^{\#*\top}\boldsymbol{x})\right]\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty}K(u)\cdot f_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}+uh)\,du\cdot\boldsymbol{x}\boldsymbol{x}^\top\right] \qquad \text{(by (EC.55))}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u)\left[f_{\epsilon|\boldsymbol{x}}(\boldsymbol{\delta}_h^{\#*\top}\boldsymbol{x}+uh)-f_{\epsilon|\boldsymbol{x}}(0)\right]du\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]+\boldsymbol{J}\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\int_{-\infty}^{\infty} K(u)\left[-2\lambda f'_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}+uh\right]du\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]+\boldsymbol{J}+o(\lambda)+o(h^2)\\
&= -2\lambda\boldsymbol{V}'+\boldsymbol{J}+o(\lambda)+o(h^2),
\end{aligned}
$$

where $\boldsymbol{V}' := \mathbb{E}_{\boldsymbol{x}}\left[f'_{\epsilon|\boldsymbol{x}}(0)(\boldsymbol{J}^{-1}\boldsymbol{\theta}^*)^\top\boldsymbol{x}\cdot\boldsymbol{x}\boldsymbol{x}^\top\right]$. Then, applying second-order approximation of an inverse matrix to $\boldsymbol{H}_h^{\#}$ leads to (EC.61).

□

EC.4.4. Proof of Proposition 3

Proof. According to Lemma EC.11, the conditional variance of non-private and private order quantities are

$$
\begin{aligned}
\mathsf{Var}\left[\widehat{q}(\boldsymbol{x})|\boldsymbol{x}\right] &= \frac{\left\|\boldsymbol{\Sigma}_r^{\frac{1}{2}}\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_2^2}{n} && \forall\boldsymbol{x}\in\mathcal{X};\\
\mathsf{Var}\left[\widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x})|\boldsymbol{x}\right] &\geq \frac{\left\|\boldsymbol{\Sigma}_r^{\frac{1}{2}}\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_2^2}{n} - c_1\lambda\frac{\left\|\boldsymbol{\Sigma}_r^{\frac{1}{2}}\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_2^2}{n} - 2h\mathsf{K}\frac{\left\|\boldsymbol{J}^{-\frac{1}{2}}\boldsymbol{x}\right\|_2^2}{n} + \frac{\sigma^2\left\|\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_2^2}{n^2}, && \forall\boldsymbol{x}\in\mathcal{X}
\end{aligned}
$$

where $\boldsymbol{J} := \mathbb{E}_{\boldsymbol{x}}\left[f_{\epsilon|\boldsymbol{x}}(0)\cdot\boldsymbol{x}\boldsymbol{x}^T\right]\succ\boldsymbol{0}$, and $\boldsymbol{\Sigma}_r := r(1-r)\boldsymbol{\Sigma}\succ\boldsymbol{0}$, $\mathsf{K} := \int_{-\infty}^{\infty}uK(u)\cdot\int_{-\infty}^{u}K(v)\,dv\cdot du\in(0,\kappa_1)$, $c_1>0$ is a constant, and $\sigma\asymp\bar{r}\sqrt{\ln(1/\delta)}/\varepsilon$. Hence, the increase in conditional variance due to DP algorithms $\Delta V(\widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x}),\widehat{q}(\boldsymbol{x})) := \mathsf{Var}\left[\widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x})|\boldsymbol{x}\right]-\mathsf{Var}\left[\widehat{q}(\boldsymbol{x})|\boldsymbol{x}\right]$ is lower bounded as

$$
\begin{aligned}
\Delta V(\widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x}),\widehat{q}(\boldsymbol{x})) &\geq \frac{(-c_1\lambda-c_2h)\cdot r(1-r)}{n}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_2^2+\frac{\sigma^2}{n^2}\left\|\boldsymbol{J}^{-1}\boldsymbol{x}\right\|_2^2, \quad \forall\boldsymbol{x}\in\mathcal{X}\\
&= \Omega\left(r(1-r)\frac{-\lambda-h}{n}+\frac{\sigma^2}{n^2}\right)\\
&= \Omega\left(r(1-r)\frac{-\lambda-h}{n}+\max\{r,1-r\}\frac{\ln(1/\delta)}{n^2\varepsilon^2}\right).
\end{aligned}
$$

The upper bound is a corollary of Proposition 2 as the increased deviation is of order σ/n.

□

EC.4.5. Formal Statement of Proposition 4 and Proof

Proposition EC.2 (Supplier Gains a Lower Profit Share; formal). *Consider a newsvendor-supplier channel under the wholesale price contract with production cost $v>0$, wholesale price $w>v$, and retail price $p>w$. Assume demand $y=\langle\boldsymbol{\theta}^\star,\boldsymbol{x}\rangle+\mathcal{N}(0,1^2)$, and Assumption 2 holds. Further assume that the optimal order quantity $q^*(\boldsymbol{x}) := \langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle$ is always positive for any $\boldsymbol{x}\in\mathcal{X}$, where $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathcal{L}(\boldsymbol{\theta})$. For any algorithm $\mathcal{A}$ associated with an estimator $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$, let the order quantity be $\widehat{q}^{\mathcal{A}}(\boldsymbol{x}) := \langle\widehat{\boldsymbol{\theta}}^{\mathcal{A}},\boldsymbol{x}\rangle,\forall\boldsymbol{x}\in\mathcal{X}$, let the expected profit of the newsvendor under policy $\mathcal{A}$ be $\widehat{\Pi}^N(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}\left[p\min\{y,\widehat{q}^{\mathcal{A}}(\boldsymbol{x})\}-w\widehat{q}^{\mathcal{A}}(\boldsymbol{x})\right],\forall\boldsymbol{x}\in\mathcal{X}$,*

and let the profit of the supplier be $\widehat{\Pi}^S(\boldsymbol{x}) = (w-v)\widehat{q}^{\mathcal{A}}(\boldsymbol{x})$*. Then, with high probability, adopting the private algorithm* OP *would raise up newsvendor's profit share (correspondingly, pull down supplier's share):*

$$\frac{\widehat{\Pi}^N(\boldsymbol{x})}{\widehat{\Pi}^N(\boldsymbol{x})+\widehat{\Pi}^S(\boldsymbol{x})} = \frac{\mathsf{Rev}^N(\boldsymbol{x})-C-U}{\mathsf{Rev}(\boldsymbol{x})-C-U-\widehat{P}(\boldsymbol{x},\mathcal{D})}, \quad \forall \boldsymbol{x}\in\mathcal{X}, \tag{EC.70}$$

by setting a positive value of $\widehat{P}(\boldsymbol{x},\mathcal{D}) := (w-v)(q^*(\boldsymbol{x}) - \widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x}))$ *in* (EC.70)*. The value* $\widehat{P}(\boldsymbol{x},\mathcal{D})$ *is the supplier's revenue loss due to newsvendor's private learning process. Other values on the right hand side of* (EC.70) *are constants:* $\mathsf{Rev}^N(\boldsymbol{x}) := (p-w)q^*(\boldsymbol{x})$ *is newsvendor's revenue at optimal order quantity;* $\mathsf{Rev}(\boldsymbol{x}) := (p-v)q^*(\boldsymbol{x})$ *is the whole chain's optimal revenue;* $C := \min_{\boldsymbol{\theta}} \mathbb{E}_{y,\boldsymbol{x}}\left[(p-w)(y-\boldsymbol{\theta}^\top\boldsymbol{x})^+ + w(\boldsymbol{\theta}^\top\boldsymbol{x}-y)^+\right]$ *is newsvendor's minimal cost;* $U := (p-w)\Phi^{-1}(\frac{p-w}{p})$ *is the revenue loss due to demand uncertainty. Infinitesimals of* $\widehat{P}(\boldsymbol{x},\mathcal{D})$ *appearing both in the numerator and denominator of the right-hand side term of* (EC.70) *are omitted.*

Proof. Let $C(\boldsymbol{\theta}) := \mathbb{E}_{y,\boldsymbol{x}}\left[(p-w)(y-\boldsymbol{\theta}^\top\boldsymbol{x})^+ + w(\boldsymbol{\theta}^\top\boldsymbol{x}-y)^+\right]$ be the newsvendor's cost at $\boldsymbol{\theta}$. By Lemma EC.8, we know that the minimizer $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})$ takes the form $\boldsymbol{\theta}^* := \boldsymbol{\theta}^\star + [\Phi^{-1}(\frac{p-w}{p}), 0, \ldots, 0]$, and the data generating process can be reformulated as $y = \boldsymbol{\theta}^{*\top}\boldsymbol{x} + \mathcal{N}(-\Phi^{-1}(\frac{p-w}{p}), 1^2)$.

Given a dataset $\mathcal{D}$, let $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ be an estimator to $\boldsymbol{\theta}^*$. Let $q^*(\boldsymbol{x}) := \boldsymbol{\theta}^{*\top}\boldsymbol{x}$ and $\widehat{q}^{\mathcal{A}}(\boldsymbol{x}) := \widehat{\boldsymbol{\theta}}^{\mathcal{A}\top}\boldsymbol{x}$ be the order quantities under context $\boldsymbol{x}$. Recall that, under context $\boldsymbol{x}$, the newsvendor's expected profit with an estimator $\widehat{\boldsymbol{\theta}}^{\mathcal{A}}$ is

$$\begin{aligned}
\widehat{\Pi}^N(\boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}}\left[p\min\{y, \widehat{\boldsymbol{\theta}}^{\mathcal{A}\top}\boldsymbol{x}\} - w\widehat{\boldsymbol{\theta}}^{\mathcal{A}\top}\boldsymbol{x}\right] \\
&= (p-w)q^*(\boldsymbol{x}) - (p-w)\Phi^{-1}\left(\frac{p-w}{p}\right) - C(\widehat{\boldsymbol{\theta}}) \\
&= (p-w)q^*(\boldsymbol{x}) - (p-w)\Phi^{-1}\left(\frac{p-w}{p}\right) - C(\boldsymbol{\theta}^*) - \mathcal{O}(\|\boldsymbol{\delta}\|_2^2),
\end{aligned}$$

where $\boldsymbol{\delta} := \widehat{\boldsymbol{\theta}}^{\mathcal{A}} - \boldsymbol{\theta}^*$, and the last equality follows from analysis by mean-value theorem:

$$\begin{aligned}
C(\widehat{\boldsymbol{\theta}}) - C(\boldsymbol{\theta}^*) &= \boldsymbol{\delta}^\top \cdot \int_0^1 \nabla C(\boldsymbol{\theta}^* + t\boldsymbol{\delta})\, dt \\
&= \boldsymbol{\delta}^\top \cdot \int_0^1 \mathbb{E}_{\boldsymbol{x}}\left[\left(\Phi\left(t\boldsymbol{\delta}^\top\boldsymbol{x} + \Phi^{-1}(\frac{p-w}{p})\right) - \frac{p-w}{p}\right)\cdot\boldsymbol{x}\right] dt \\
&= \mathcal{O}(\|\boldsymbol{\delta}\|_2^2) > 0.
\end{aligned}$$

Similarly, under context $\boldsymbol{x}$, the supplier's profit is

$$\widehat{\Pi}^S(\boldsymbol{x}) = (w-v)\widehat{q}^{\mathcal{A}}(\boldsymbol{x}).$$

Therefore, when the newsvendor adopts Algorithm OP, the profit share of the newsvendor is:

$$\frac{\widehat{\Pi}^N(\boldsymbol{x})}{\widehat{\Pi}^N(\boldsymbol{x})+\widehat{\Pi}^S(\boldsymbol{x})} = \frac{(p-w)q^*(\boldsymbol{x}) - (p-w)\Phi^{-1}\left(\frac{p-w}{p}\right) - C(\boldsymbol{\theta}^*) - \mathcal{O}(\|\boldsymbol{\delta}\|_2^2)}{(p-w)q^*(\boldsymbol{x}) - (p-w)\Phi^{-1}\left(\frac{p-w}{p}\right) - C(\boldsymbol{\theta}^*) - \mathcal{O}(\|\boldsymbol{\delta}\|_2^2) + (w-v)\widehat{q}(\boldsymbol{x})}$$

$$= \frac{\mathsf{Rev}^N(\boldsymbol{x}) - U - C - \mathcal{O}(\|\boldsymbol{\delta}\|_2^2)}{\mathsf{Rev}(\boldsymbol{x}) - U - C - \widehat{P}(\boldsymbol{x},\mathcal{D}) + \mathcal{O}(\|\boldsymbol{\delta}\|_2^2)}$$
$$= \frac{\mathsf{Rev}^N(\boldsymbol{x}) - U - C - o(\widehat{P}(\boldsymbol{x},\mathcal{D}))}{\mathsf{Rev}(\boldsymbol{x}) - U - C - \widehat{P}(\boldsymbol{x},\mathcal{D}) - o(\widehat{P}(\boldsymbol{x},\mathcal{D}))}.$$

The second line basically only involves notation changes: $\mathsf{Rev}^N(\boldsymbol{x}) := (p-w)q^*(\boldsymbol{x})$, $\widehat{P}(\boldsymbol{x},\mathcal{D}) := (w-v)(q^*(\boldsymbol{x}) - \widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x}))$; $U := (p-w)\Phi^{-1}(\frac{p-w}{p})$; $C := C(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta}} \mathbb{E}_{y,\boldsymbol{x}}\left[(p-w)(y-\boldsymbol{\theta}^\top\boldsymbol{x})^+ + w(\boldsymbol{\theta}^\top\boldsymbol{x}-y)^+\right]$; and $\mathsf{Rev}(\boldsymbol{x}) := (p-v)q^*(\boldsymbol{x})$. In the last line, we replace $\mathcal{O}(\|\boldsymbol{\delta}\|_2^2)$ with $o(\widehat{P}(\boldsymbol{x},\mathcal{D}))$ since $\widehat{P}(\boldsymbol{x},\mathcal{D}) = \mathcal{O}(\|\boldsymbol{\delta}\|_2)$.

To show the profit share of the supplier is decreased, it suffices to show $\widehat{P}(\boldsymbol{x},\mathcal{D}) := (w-v)(q^*(\boldsymbol{x}) - \widehat{q}_h^{\mathsf{OP}}(\boldsymbol{x}))$ is positive with high probability. Recall Proposition 2, we know that

$$\widehat{P}(\boldsymbol{x},\mathcal{D}) = (w-v)\left[2\lambda\boldsymbol{u}^\top\boldsymbol{x} + \frac{1}{2}h^2\kappa_2\boldsymbol{v}^\top\boldsymbol{x}\right] + \mathcal{O}_p\left(\frac{1}{\sqrt{n}} + \frac{\sigma}{n}\right),$$

with $\boldsymbol{u} = \boldsymbol{J}^{-1}\boldsymbol{\theta}^*$, $\boldsymbol{v} = \boldsymbol{J}^{-1}\mathbb{E}_{\boldsymbol{x}}\left[f_\epsilon'(0)\boldsymbol{x}\right]$, and $\boldsymbol{J} := \mathbb{E}_{\boldsymbol{x}}\left[f_\epsilon(0)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right] \succ \boldsymbol{0}$. As we assume $q^*(\boldsymbol{x}) := \boldsymbol{\theta}^{*\top}\boldsymbol{x}$ is always positive, the value $\boldsymbol{u}^\top\boldsymbol{x}$ is therefore always positive. Additionally, under our assumptions (Assumption 2, standardized $\boldsymbol{x}$), we have $\boldsymbol{v}^\top\boldsymbol{x} = f_\epsilon'(0)/f_\epsilon(0)$. Since $\epsilon \sim \mathcal{N}(-\Phi^{-1}(\frac{p-w}{p}), 1)$:

- when $(p-w)/p \le 1/2$ (i.e., $w \ge p/2$), we have $f_\epsilon'(0) \ge 0$, which ensures $\boldsymbol{v}^\top\boldsymbol{x}$ is always non-negative, and further ensures $\widehat{P}(\boldsymbol{x},\mathcal{D})$ is positive w.h.p;
- when $(p-w)/p > 1/2$ (i.e., $w < p/2$), $\boldsymbol{v}^\top\boldsymbol{x}$ is negative. Nevertheless, recall that when running OP, we require $\lambda \gtrsim h^2$ (this can be verified by the values of λ, h, ε, and analysis around Eq. (EC.19)). Hence, the positive value of $\boldsymbol{u}^\top\boldsymbol{x}$ will finally dominates the negative value of $\boldsymbol{v}^\top\boldsymbol{x}$, resulting in a positive $\widehat{P}(\boldsymbol{x},\mathcal{D})$ w.h.p.

Combining both cases completes the proof.

□

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318. 14

Agarwal N, Kale S, Singh K, Thakurta A (2023) Differentially private and lazy online convex optimization. *The Thirty Sixth Annual Conference on Learning Theory*, 4599–4632 (PMLR).

Asi H, Feldman V, Koren T, Talwar K (2021) Private stochastic convex optimization: Optimal rates in l1 geometry. *International Conference on Machine Learning*, 393–403 (PMLR). 8, 11

Beck A, Teboulle M (2012) Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization* 22(2):557–580.

Billingsley P (2017) *Probability and measure* (John Wiley & Sons).

Bousquet O, Elisseeff A (2002) Stability and generalization. *The Journal of Machine Learning Research* 2:499–526. 11

Cai TT, Wang Y, Zhang L (2021) The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics* 49(5):2825–2850.

Cai TT, Wang Y, Zhang L (2023) Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152* .

Dong J, Roth A, Su WJ (2022) Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84(1):3–37. 13

Duchi JC, Bartlett PL, Wainwright MJ (2012) Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization* 22(2):674–701.

Dwork C, Rothblum GN, Vadhan S (2010) Boosting and differential privacy. *2010 IEEE 51st annual symposium on foundations of computer science*, 51–60 (IEEE).

Fernandes M, Guerre E, Horta E (2021) Smoothing quantile regressions. *Journal of Business & Economic Statistics* 39(1):338–357. 8, 17

Hardt M, Recht B, Singer Y (2016) Train faster, generalize better: Stability of stochastic gradient descent. *International conference on machine learning*, 1225–1234 (PMLR).

Kifer D, Smith A, Thakurta A (2012) Private convex empirical risk minimization and high-dimensional regression. *Conference on Learning Theory*, 25–1 (JMLR Workshop and Conference Proceedings). 10, 11, 12, 21, ec5, ec23

Mironov I (2017) Rényi differential privacy. *2017 IEEE 30th computer security foundations symposium (CSF)*, 263–275 (IEEE).

Redberg R, Koskela A, Wang YX (2024) Improving the privacy and practicality of objective perturbation for differentially private linear learners. *Advances in Neural Information Processing Systems* 36.

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

Vadhan S (2017) The complexity of differential privacy. *Tutorials on the Foundations of Cryptography*, 347–450 (Springer).

Van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge university press).

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