
Algorithm 3: Mixed-Sample SGD (Restated for General Θ)

Input: $\theta_0 = \theta_{Q,0} = 0$, $\lambda_0 = 0$, stepsize $\{\alpha_t\}_{t=0}^{T-1}$ and η, γ, ϵ_Q .

for $t = 0, \dots, T-1$ **do**

 Draw $\xi_t \sim \text{Bernoulli}(\frac{1}{1+\lambda_t})$

if $\xi_t = 1$ **then**

 Sample (x_t, y_t) uniformly from S_P

$\theta_{t+1} = \mathcal{P}_\Theta(\theta_t - \eta(1 + \lambda_t)\nabla\ell(\theta_t; x_t, y_t))$

end

else

 Sample (x_t, y_t) uniformly from S_Q

$\theta_{t+1} = \mathcal{P}_\Theta(\theta_t - \eta(1 + \lambda_t)\nabla\ell(\theta_t; x_t, y_t))$

end

 Sample (x_t, y_t) uniformly from S_Q

$\lambda_{t+1} = [(1 - \gamma\eta)\lambda_t + \eta(\ell(\theta_t; x_t, y_t) - \ell(\theta_{Q,t}; x_t, y_t) - 3\epsilon_Q)]_+$

$\theta_{Q,t+1} = \mathcal{P}_\Theta(\theta_{Q,t} - \alpha_t\nabla\ell(\theta_{Q,t}; x_t, y_t))$

end

$\hat{\theta}_{PQ} = \ell_2$ projection of $\frac{1}{T} \sum_{t=0}^{T-1} \theta_t$ onto the constraint set $\{\theta : \hat{R}_Q(\theta) - \hat{R}_Q(\theta_{Q,T}) \leq 2\epsilon_Q\}$.

Output: $\hat{\theta}_{PQ}$

360 8 Results for General Losses

361 In this section, we provide the results for general convex losses. We consider a hypothesis class
362 $\mathcal{H} \doteq \{x \mapsto h_\theta(x) : \theta \in \Theta\}$. We make the following assumption on the loss function ℓ .

363 **Assumption 3.** $\ell(\theta; x, y)$ is m_1 -strongly convex and m_2 -smooth in θ for any $(x, y) \in S_{PQ}$.

364 We define $\kappa \doteq m_1/m_2$ as the condition number of ℓ .

365 We consider general $\Theta \subseteq \mathbb{R}^D$, possibly a strict subset, and therefore present a version of the algorithm
366 that projects iterates back to Θ , provided a projection operator \mathcal{P}_Θ (see Algorithm 3); when $\Theta = \mathbb{R}^D$,
367 i.e. is unbounded, the operator reduces to identity mapping so Algorithm 3 reduces to Algorithm 1 in
368 the main paper. Note that this projection operator is usually cheap for the common choices of Θ : for
369 example, for the sake of regularization in practice (e.g., ridge-type regularization), often Θ is an ℓ_2
370 unit ball, whereby $\mathcal{P}_\Theta(\theta) = \theta / \max\{1, \|\theta\|\}$.

371 For the convergence analysis, we need the following definitions.

372 **Definition 4** (Key Lipschitz Parameters). Let $\rho \doteq \|\theta_0 - \tilde{\theta}_{PQ}\|$. We then define

$$\begin{aligned} \hat{G}_\theta &= \sup \left\{ \|\nabla\ell(\theta; x, y)\| : \|\theta - \tilde{\theta}_{PQ}\|^2 \leq 2\rho^2, (x, y) \in S_{PQ}, \right. \\ &\quad \left. \|\theta\| \leq \frac{1 + \log(T + \kappa)}{m_1} \sup_{(x', y') \in S_Q} \|\nabla\ell(\theta_{Q,0}; x', y')\| \right\}, \\ \hat{G}_\lambda &= \sup \left\{ |\ell(\theta; x, y) - \ell(\theta'; x, y) - 3\epsilon_Q| : \|\theta - \tilde{\theta}_{PQ}\|^2 \leq 2\rho^2, (x, y) \in S_{PQ}, \right. \\ &\quad \left. \|\theta'\| \leq \frac{1 + \log(T + \kappa)}{m_1} \sup_{(x', y') \in S_Q} \|\nabla\ell(\theta_{Q,0}; x', y')\| \right\}. \end{aligned}$$

373 The above definition captures the Lipschitzness of the risk function on the iterates generated during
374 our algorithm proceeding.

375 **Definition 5** (See [6]). For $\epsilon > 0$, let $r(\epsilon) \doteq \inf \left\{ \|\nabla\hat{R}_Q(\theta)\|^2 : \theta \in \Theta, \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) = \epsilon \right\}$.

376 This notion is used to control the distance between the return solution and solution before the last
377 projection step, i.e., $\|\theta_T - \hat{\theta}_{PQ}\|$ where $\theta_T \doteq \frac{1}{T} \sum_{t=0}^{T-1} \theta_t$. In linear regression, we can explicitly
378 compute it as $r(\epsilon) = \epsilon \cdot \lambda_{\min}^+(\hat{\Sigma}_Q)$.

Theorem 4 (Strongly Convex and Smooth Result). *Let $\rho^2 = \|\theta_0 - \tilde{\theta}_{PQ}\|^2$, $\tau \leq 0.1$, and σ_{PQ}^2, c_η be some positive numbers such that*

$$\sigma_{PQ}^2 \leq 256 (1 + (\lambda^*)^2 + \rho^2) \hat{G}_\theta^2, \text{ and } c_\eta \leq \min \left\{ \frac{\rho}{2\sqrt{2}\hat{G}_\lambda}, \frac{\rho}{16\sqrt{6}\sigma_{PQ}\sqrt{\log 2/\tau}}, \frac{\rho}{C_{PQ}} \right\},$$

where $C_{PQ} = 2m_1(1 + 2\kappa) \left(\frac{1}{m_1^2} \|\nabla \hat{R}_Q(\theta_{Q,0})\|^2 + \|\theta_{Q,0}\|^2 \right) + \frac{6\sigma_Q^2 \log(2/\tau)}{m_1}$, for $\sigma_Q^2 = \left(\frac{\log(T+2\kappa)}{m_1} \sup_{(x,y) \in S_Q} \|\nabla \ell(\theta_{Q,0}; x, y)\| \right)^2$.

Assume the parameters in Algorithm 3 are set as $\eta = \frac{c_\eta}{\sqrt{T}}$, $\gamma = \frac{\hat{G}_\theta c_\eta}{\sqrt{T}}$, $\alpha_t = \frac{1}{m_1} \cdot \frac{1}{t+2\kappa}$. Suppose $T \geq \max \left\{ \frac{C_{PQ} \log T}{\epsilon_Q}, \left(\frac{C_{PQ} (\log(T+2\kappa))^2}{\hat{G}_\lambda \sqrt{\log 1/\tau}} \right)^2 \right\}$ and $r(2\epsilon_Q) \leq 1$, then, with probability at least $1 - \tau$ over the randomness of Algorithm 3, the empirical risk of the returned solution satisfies:

$$\hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \lesssim \left(\hat{G}_\theta + \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}} \right) \cdot \left(\frac{\frac{\hat{G}_\theta^2}{c_\eta} + \hat{G}_\theta \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}}}{r(2\epsilon_Q) \sqrt{T}} + \frac{\lambda^* \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}} + \frac{\rho^2}{c_\eta}}{\sqrt{T}} \right).$$

Statistical Implications. In this section we give an example of how the above Theorem 4 can be converted generically to statistical guarantees on transfer (as was done for the linear regression case).

For intuition, Theorem 4 guarantees that $\hat{R}_P(\hat{\theta}_{PQ})$ is small, i.e., close to $\hat{R}_P(\tilde{\theta}_{PQ})$, while we also know $\hat{R}_Q(\hat{\theta}_{PQ})$ is also small, i.e., close to $\hat{R}_Q(\hat{\theta}_Q)$ (since $\hat{R}_P(\hat{\theta}_{PQ})$ is a projection of the average iterate onto the constraint set centered at $\hat{\theta}_{Q,T}$). Thus, if in addition, we have concentration of empirical risks to true risks, we can convert the guarantees of Theorem 4 to statistical transfer guarantees.

The results below illustrate the above intuition. These results are expressed in terms of generic relations between P and Q risks given as follows.

Definition 6 (Weak Modulus [5]). *Given $\epsilon > 0$, we define the modulus*

$$\delta(\epsilon) \doteq \sup \{ \mathcal{E}_Q(\theta) : \mathcal{E}_P(\theta) \leq \epsilon, \theta \in \Theta \}.$$

In words, the weak modulus captures the best achievable Q risk, if the learner only has access to P 's data. For instance, in linear regression, as explained in the main paper and shown in [5] it can be upper bounded as $\delta(\epsilon) \leq 2\lambda_{\max}(\Sigma_P^{-1} \Sigma_Q) \cdot \epsilon + 2\mathcal{E}_Q(\theta_P^*)$.

We next consider some traditional concentration results for bounded losses in terms of the Rademacher complexity of the loss class.

Assumption 4 (Boundedness of Loss). *We assume $\ell(\theta; x, y) \leq M_\ell$ for any $\theta \in \Theta$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$.*

Remark 3. *The Assumption 4 hold for a strongly convex loss given that the parameter θ 's norm is bounded, e.g., $\|\theta\| \leq B$, $\forall \theta \in \Theta$. For example, in linear regression, if we assume the label space $\mathcal{Y} \subseteq [-M_y, M_y]$, then $\ell(\theta; x, y) = (\theta^\top x - y)^2 \leq 2B^2 M_x^2 + 2M_y^2$.*

We then introduce the Rademacher complexity which characterizes the complexity of a class and will be used to derive uniform convergence result.

Definition 7 (Rademacher complexity). *Let \mathcal{F} be a family of functions mapping from \mathcal{Z} to \mathbb{R} and $Z = \{z_1, \dots, z_n\}$ be the i.i.d. samples drawn from distribution μ . Then, the empirical Rademacher complexity of \mathcal{F} with respect to dataset Z is defined as*

$$\hat{\mathcal{R}}_n(\mathcal{F}) \doteq \mathbb{E}_{\epsilon \in \{\pm 1\}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(z_i) \right],$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Rademacher random variables with $\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$. Then Rademacher complexity $R_n(\mathcal{F})$ is defined as $R_n(\mathcal{F}) \doteq \mathbb{E} \hat{\mathcal{R}}_n(\mathcal{F})$

Assumption 5 (Bounded Rademacher Complexity of Loss Class). *We assume $\mathcal{R}_n(\ell \circ \mathcal{H}) \leq \frac{B_{\mathcal{H}, \ell}}{\sqrt{n}}$ for some positive real number $B_{\mathcal{H}, \ell}$ which characterizes the complexity of the loss class $\ell \circ \mathcal{H}$.*

414 **Remark 4.** The Assumption 5 is standard. Here we give some examples.

415 For linear classifier class $\mathcal{H} \doteq \{x \mapsto \theta^\top x : \theta \in \mathbb{R}^d, \|\theta\| \leq B\}$ with L -Lipschitz loss, the bound [23,
416 Lemma 26.10] is given by $\mathcal{R}_n(\ell \circ \mathcal{H}) \leq \frac{LBM_x}{\sqrt{n}}$.

417 For l layer neural network class $\mathcal{H} \doteq \{x \mapsto W_l \sigma(W_{l-1} \dots \sigma(W_1 x)) : \|W_j\|_F \leq B_j\}$ with L -
418 Lipschitz loss, the bound [24, Theorem 1] is given by

$$\mathcal{R}_n(\ell \circ \mathcal{H}) \lesssim \frac{LM_x \sqrt{l} \prod_{j=1}^l B_j}{\sqrt{n}}.$$

419 For general VC class $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$ with VC dimension d , the bound [25, Corollary 3.8] is given by
420 $\mathcal{R}_n(\ell \circ \mathcal{H}) \lesssim \sqrt{\frac{d \log n}{n}}$.

421 With the above two assumptions we can derive the following rate of uniform convergence.

422 **Proposition 1** (Uniform Convergence). *Let μ denote either P or Q . With probability at least $1 - \tau$,*
423 *the following statement holds:*

$$\sup_{h \in \mathcal{H}} |R_\mu(\theta) - \hat{R}_\mu(\theta)| \leq 2 \frac{B_{\mathcal{H}, \ell}}{\sqrt{n_\mu}} + M_\ell \sqrt{\frac{\log \frac{2}{\tau}}{2n_\mu}}.$$

Corollary 1 (Statistical Transfer Guarantees). *Let*

$$\epsilon_P = 2 \frac{B_{\mathcal{H}, \ell}}{\sqrt{n_P}} + M_\ell \sqrt{\frac{\log \frac{2}{\tau}}{2n_P}}, \epsilon_Q = 2 \frac{B_{\mathcal{H}, \ell}}{\sqrt{n_Q}} + M_\ell \sqrt{\frac{\log \frac{2}{\tau}}{2n_Q}}.$$

424 *Then with probability at least $1 - 3\tau$ over the randomness of Algorithm 3 and S_P and S_Q , the*
425 *returned solution satisfies*

$$\mathcal{E}_Q(\hat{\theta}_{PQ}) \leq 5 \cdot \min \{\epsilon_Q, \delta(3\epsilon_P)\},$$

426 *provided a number of iterations*

$$T \gtrsim \left(\hat{G}_\theta + \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}} \right)^2 \cdot \left(\frac{\frac{\hat{G}_\theta^2}{c_\eta} + \hat{G}_\theta \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}}}{r(2\epsilon_Q)\epsilon_P} + \frac{\lambda^* \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}} + \frac{\rho^2}{c_\eta}}{\epsilon_P} \right)^2.$$

427 Here we achieve a transfer guarantee similar to that in Theorem 1. The rate is still adaptive—it
428 achieves the better rate between solely target ERM and source ERM. The required iteration number
429 depends on the ϵ_P and $r(\epsilon_Q)$, also similar to that in Theorem 1.

430

431 9 Additional Experiments

432 In this section we provide additional experimental results.

433 9.1 Regression Task on the Berkeley Yearbook Dataset.

434 We conduct experiments on the Berkeley Yearbook Dataset [26]. The dataset contains the gray-scale
435 portraits taken in different years. The input x is the 512-dimensional vector feature extracted by
436 ResNet18, and y is the year the photo is taken (ranging from 1905 to 2013). We construct source
437 and target tasks by varying the proportion of male and female photos. In the source dataset, 50% of
438 the samples are drawn from male photos and 50% from female photos. For the target training and
439 testing dataset, the ratio is adjusted to 75% male and 25% female. We fix $n_Q = 100$ and vary n_P
440 from 500 to 1300. Due to the difficulty of the task and the limited target data, the target ERM model
441 suffers from very large errors and is therefore omitted from Figure 5. We report the MSE comparison
442 with source ERM and HTL, as well as a runtime comparison with HTL. We can see from the left

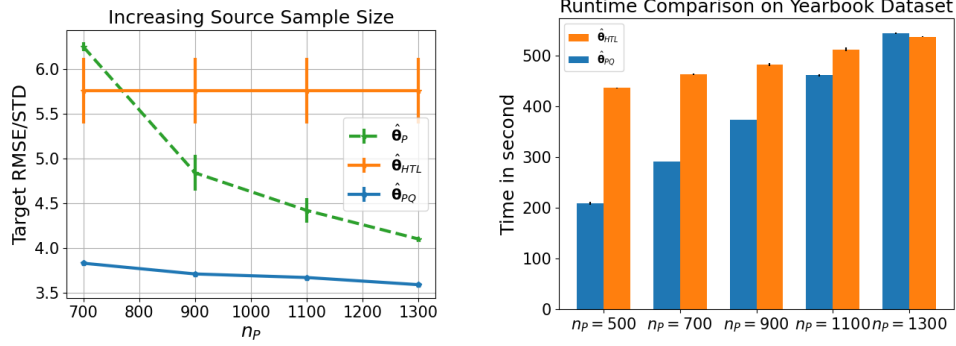


Figure 5: Linear regression on the Yearbook dataset. Input features of dimension 512 are extracted by ResNet18. (Left) Q excess risk. (Right) Runtime comparison. HTL ($\hat{\theta}_{HTL}$) has difficulty to adapt to the situation where source data are more helpful and results in large target error, while our algorithm ($\hat{\theta}_{PQ}$) gains from source data and significantly outperforms the competitors.

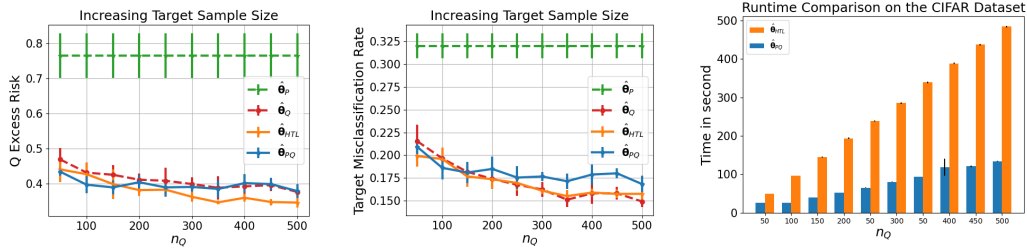


Figure 6: Binary Classification on the CIFAR10 Dog vs Cat dataset. We fix $n_P = 100$ and increase n_Q gradually. (Left) Target excess risk (Middle) Target misclassification rate. (Right) Runtime comparison. Our method ($\hat{\theta}_{PQ}$) is still comparable with target ERM. HTL ($\hat{\theta}_{HTL}$) slightly outperforms ours, but as n_Q increases, the runtime of HTL dramatically increases.

sub-figure that our method can consistently outperform source ERM, target ERM and HTL, and can automatically adapt to the better rate of source and target ERM learning. The right sub-figure shows that our algorithm achieves superior runtime performance compared to HTL when $n_P < 1300$, while when n_P reaches 1300, HTL becomes the faster one. This difference arises from the inherent nature of the two algorithms: our method primarily optimizes over the source data, so the total number of iterations increases with n_P grows, In contrast, HTL focuses on optimizing over the target data, using the source data only to compute a reference model. As a result, its runtime remains relatively stable as n_P increases.

9.2 More Results on CIFAR10 Dog vs Cat dataset

Here we provide more results on CIFAR10. To verify the adaptivity of the algorithm with increasing target samples, we conduct the experiment with fixed $n_P = 100$ and gradually increasing n_Q from 50 to 500. We set the source dataset to be 80% dog samples and 20% cat samples, and the ratio for target is adjusted to 50% dog and 50% cat. As we can see from Figure 6, in this setting source data are not informative so the source ERM performs poorly. As the target sample size increases, the target ERM can give promising performance, and our method can also adapt to it. HTL performs the best in this setting, but its runtime dramatically increases as the target sample size increases.

10 Missing Proofs in Section 4.1

In this section, we provide the missing proofs in Section 4.1. We first introduce the following helper lemma that lower bounds the norm of the constraint gradient on the boundard of the constraint set.

462 **Lemma 8** (Lower and upper bound of boundary gradient). *For any $\epsilon > 0$, the following statements*
 463 *hold:*

$$\min_{\theta: \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) = \epsilon} \left\| \nabla \hat{R}_Q(\theta) \right\|^2 = \epsilon \lambda_{\min}^+(\hat{\Sigma}_Q).$$

464 *Proof.* We start by proving the first statement. First, the θ on the boundary of the constraint set
 465 satisfies: $\left\| \theta - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 = \epsilon$. We first decompose θ as $\theta = \theta' + \theta^\perp$ where $\theta' \in \text{Range}(\hat{\Sigma}_Q)$ and θ^\perp is
 466 in the null space of $\hat{\Sigma}_Q$.

467 Now we examine the gradient norm:

$$\begin{aligned} \min_{\theta: \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) = \epsilon} \left\| \nabla \hat{R}_Q(\theta) \right\|^2 &= \min_{\left\| \theta - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 = \epsilon} \left\| \hat{\Sigma}_Q(\theta - \hat{\theta}_Q) \right\|^2 \\ &= \min_{\left\| \theta' - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 = \epsilon, \theta' \in \text{Range}(\hat{\Sigma}_Q)} \left\| \hat{\Sigma}_Q(\theta' - \hat{\theta}_Q) \right\|^2 \\ &= \min_{\mathbf{u} \in \text{Range}(\hat{\Sigma}_Q), \|\mathbf{u}\|=1} \left\| \sqrt{\epsilon} \hat{\Sigma}_Q^{\frac{1}{2}} \mathbf{u} \right\|^2 = \epsilon \lambda_{\min}^+(\hat{\Sigma}_Q). \end{aligned}$$

468 □

469 **Proof of Lemma 1:**

470 *Proof.* Due to Jensen's inequality, we have that $\left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2 \leq 2 \left\| \theta_0 - \hat{\theta}_P \right\|^2 + 2 \left\| \hat{\theta}_P - \tilde{\theta}_{PQ} \right\|^2$,
 471 which is at most

$$\frac{1}{2\lambda_{\min}^2(\hat{\Sigma}_P)} \left\| \nabla \hat{R}_P(\theta_0) \right\|^2 + 2 \left\| \hat{\theta}_P - \tilde{\theta}_{PQ} \right\|^2 \leq \frac{\left\| \nabla \hat{R}_P(\theta_0) \right\|^2}{2\lambda_{\min}^2(\hat{\Sigma}_P)} + \frac{2 \left\| \hat{\theta}_P - \tilde{\theta}_{PQ} \right\|_{\hat{\Sigma}_P}^2}{\lambda_{\min}(\hat{\Sigma}_P)}.$$

472 Since $\tilde{\theta}_{PQ}$ is the minimizer of \hat{R}_P within constraint set, and $\hat{\theta}_Q$ is also in the constraint set, we have

$$\left\| \hat{\theta}_P - \tilde{\theta}_{PQ} \right\|_{\hat{\Sigma}_P}^2 \leq \left\| \hat{\theta}_P - \hat{\theta}_Q \right\|_{\hat{\Sigma}_P}^2 \leq \frac{\left\| \nabla \hat{R}_P(\hat{\theta}_Q) \right\|^2}{4\lambda_{\min}(\hat{\Sigma}_P)}.$$

473 Putting pieces together we have

$$\left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2 \leq \frac{\left\| \nabla \hat{R}_P(\theta_0) \right\|^2}{2\lambda_{\min}^2(\hat{\Sigma}_P)} + \frac{\left\| \nabla \hat{R}_P(\hat{\theta}_Q) \right\|^2}{2\lambda_{\min}^2(\hat{\Sigma}_P)}.$$

474 □

475 **Proof of Lemma 2:**

476 *Proof.* For θ such that $\left\| \theta - \tilde{\theta}_{PQ} \right\|^2 \leq 2\rho^2$, we examine the upper bound of its norm. According to
 477 triangle inequality we have:

$$\left\| \theta \right\| \leq \sqrt{2\rho} + \left\| \tilde{\theta}_{PQ} \right\| \leq 3\rho.$$

478 The rest of the proof follows:

$$\left\| \nabla \ell(\theta; x, y) \right\| = \left\| x(\theta^\top x - y) \right\| \leq M_x^2 \left\| \theta \right\| + M_x \hat{M}_y.$$

479 □

480 **Proof of Lemma 3:**

481 *Proof.* The proof simply follows the definition of ℓ :

$$\begin{aligned}
|\ell(\theta; x, y) - \ell(\theta'; x, y) - 6\epsilon_Q| &\leq (\theta^\top x - y)^2 + (\theta'^\top x - y)^2 + 6\epsilon_Q \\
&\leq 2M_x^2 \|\theta\|^2 + 2\hat{M}_y + 2M_x^2 \|\theta'\|^2 + \epsilon_Q \\
&\leq 18 \left(M_x^2 \rho^2 + \hat{M}_y + \frac{M_x^4 \hat{M}_y^2 (1 + \log(T + 2\kappa_Q))^2}{(\lambda_{\min}^+(\hat{\Sigma}_Q))^2} \right) + 6\epsilon_Q.
\end{aligned}$$

482

□

483 **Proof of Lemma 4:**

484 *Proof.* Due to the first order optimality condition we have

$$\nabla \hat{R}_P(\tilde{\theta}_{PQ}) + \lambda^* \nabla \hat{R}_Q(\tilde{\theta}_{PQ}) = 0 \implies \lambda^* = \frac{\|\nabla \hat{R}_P(\tilde{\theta}_{PQ})\|}{\|\nabla \hat{R}_Q(\tilde{\theta}_{PQ})\|}.$$

485 To upper bound $\|\nabla \hat{R}_P(\tilde{\theta}_{PQ})\|$ we notice that

$$\begin{aligned}
\|\nabla \hat{R}_P(\tilde{\theta}_{PQ})\| &= \|2\hat{\Sigma}_P(\tilde{\theta}_{PQ} - \hat{\theta}_P)\| \\
&\leq 2\sqrt{\lambda_{\max}(\hat{\Sigma}_P)} \|\hat{\Sigma}_P^{1/2}(\tilde{\theta}_{PQ} - \hat{\theta}_P)\| = 2\sqrt{\lambda_{\max}(\hat{\Sigma}_P)} \sqrt{\hat{R}_P(\tilde{\theta}_{PQ}) - \hat{R}_P(\hat{\theta}_P)}.
\end{aligned}$$

486 Define θ' to be the model on the boundary of constraint set, and also on the range of $\hat{\Sigma}_Q$. That is,

487 $\|\theta' - \hat{\theta}_Q\|_{\hat{\Sigma}_Q}^2 = 6\epsilon_Q$ and $\theta' \in \text{Range}(\hat{\Sigma}_Q)$. Since $\tilde{\theta}_{PQ}$ is the minimizer of $\hat{R}_P(\cdot)$ in the constraint
488 set, we know that

$$\begin{aligned}
\|\nabla \hat{R}_P(\tilde{\theta}_{PQ})\| &\leq 2\sqrt{\lambda_{\max}(\hat{\Sigma}_P)} \sqrt{\hat{R}_P(\theta') - \hat{R}_P(\hat{\theta}_P)} \\
&= 2\sqrt{\lambda_{\max}(\hat{\Sigma}_P)} \|\hat{\Sigma}_P^{1/2}(\theta' - \hat{\theta}_P)\| \\
&\leq 2\lambda_{\max}(\hat{\Sigma}_P) \|\theta' - \hat{\theta}_P\| \\
&\leq 2\lambda_{\max}(\hat{\Sigma}_P) \|\theta' - \hat{\theta}_Q\| + 2\lambda_{\max}(\hat{\Sigma}_P) \|\hat{\theta}_Q - \hat{\theta}_P\| \\
&\leq 2\lambda_{\max}(\hat{\Sigma}_P) \frac{1}{\sqrt{\lambda_{\min}^+(\hat{\Sigma}_Q)}} \|\hat{\Sigma}_Q^{1/2}(\theta' - \hat{\theta}_Q)\| + 2\lambda_{\max}(\hat{\Sigma}_P) \|\hat{\theta}_Q - \hat{\theta}_P\| \\
&= \frac{2\lambda_{\max}(\hat{\Sigma}_P)}{\sqrt{\lambda_{\min}^+(\hat{\Sigma}_Q)}} \cdot \sqrt{6\epsilon_Q} + 2\lambda_{\max}(\hat{\Sigma}_P) \|\hat{\theta}_Q - \hat{\theta}_P\|.
\end{aligned}$$

489 To lower bound $\|\nabla \hat{R}_Q(\tilde{\theta}_{PQ})\|$, we again evoke Lemma 8 that $\|\nabla \hat{R}_Q(\tilde{\theta}_{PQ})\| \geq 2\sqrt{\lambda_{\min}^+(\hat{\Sigma}_Q)} \cdot 6\epsilon_Q$.

490 Putting pieces together will conclude the proof.

491

□

492 **11 Missing Proofs in Section 5.2**

493 **Proof of Lemma 5:**

494 *Proof.* The proof mainly follows Proposition 8 of [5]. With probability at least $1 - 2\tau$, the following
 495 two facts hold. For one hand, for any $\theta \in \left\{ \theta : \left\| \theta - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 \leq 6\epsilon_Q \right\}$, we know

$$\mathcal{E}_Q(\theta) = \left\| \theta - \theta_Q^* \right\|_{\Sigma_Q}^2 \leq 2 \left\| \theta - \hat{\theta}_Q \right\|_{\Sigma_Q}^2 + 2 \left\| \theta_Q^* - \hat{\theta}_Q \right\|_{\Sigma_Q}^2 \leq 4 \left\| \theta - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 + 2\epsilon_Q \leq 26\epsilon_Q.$$

496 where at the second inequality we evoke the matrix concentration result from Lemma 3 of [5]. For
 497 the other hand, for any θ such that $\mathcal{E}_Q(\theta) \leq \epsilon_Q$, we have

$$\left\| \theta - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 \leq \frac{3}{2} \left\| \theta - \hat{\theta}_Q \right\|_{\Sigma_Q}^2 \leq 3 \left\| \theta - \theta_Q^* \right\|_{\Sigma_Q}^2 + 3 \left\| \hat{\theta}_Q - \theta_Q^* \right\|_{\Sigma_Q}^2 \leq 6\epsilon_Q.$$

498 □

499 **Proof of Lemma 6:**

500 *Proof.* First notice the following decomposition: $\left\| \theta - \theta_P^* \right\|_{\Sigma_P}^2 \leq 2 \left\| \theta - \tilde{\theta}_{PQ} \right\|_{\Sigma_P}^2 +$
 501 $2 \left\| \tilde{\theta}_{PQ} - \theta_P^* \right\|_{\Sigma_P}^2$. For the first term in RHS of above inequality, with probability at least $1 - \tau$, we
 502 have

$$2 \left\| \theta - \tilde{\theta}_{PQ} \right\|_{\Sigma_P}^2 \leq 4 \left\| \theta - \tilde{\theta}_{PQ} \right\|_{\hat{\Sigma}_P}^2 = 4 \left(\hat{R}_P(\theta) - \hat{R}_P(\tilde{\theta}_{PQ}) - \left\langle \nabla \hat{R}_P(\tilde{\theta}_{PQ}), \theta - \tilde{\theta}_{PQ} \right\rangle \right).$$

503 Since both θ and $\tilde{\theta}_{PQ}$ are in the constraint set of Problem (2), and $\tilde{\theta}_{PQ}$ is the optimal solution
 504 within the set, we know $\left\langle \nabla \hat{R}_P(\tilde{\theta}_{PQ}), \theta - \tilde{\theta}_{PQ} \right\rangle \geq 0$. Hence, we know $2 \left\| \theta - \tilde{\theta}_{PQ} \right\|_{\Sigma_P}^2 \leq$
 505 $4 \left(\hat{R}_P(\theta) - \hat{R}_P(\tilde{\theta}_{PQ}) \right)$.

506 Now we proceed to bounding $2 \left\| \tilde{\theta}_{PQ} - \theta_P^* \right\|_{\Sigma_P}^2$. Notice that with probability at least $1 - 2\tau$

$$\begin{aligned} 2 \left\| \tilde{\theta}_{PQ} - \theta_P^* \right\|_{\Sigma_P}^2 &\leq 4 \left\| \tilde{\theta}_{PQ} - \hat{\theta}_P \right\|_{\Sigma_P}^2 + 4 \left\| \hat{\theta}_P - \theta_P^* \right\|_{\Sigma_P}^2 \\ &\leq 8 \left\| \tilde{\theta}_{PQ} - \hat{\theta}_P \right\|_{\hat{\Sigma}_P}^2 + 4 \left\| \hat{\theta}_P - \theta_P^* \right\|_{\Sigma_P}^2 \\ &\leq 8 \left\| \theta_P^* - \hat{\theta}_P \right\|_{\hat{\Sigma}_P}^2 + 4 \left\| \hat{\theta}_P - \theta_P^* \right\|_{\Sigma_P}^2 \\ &\leq 12 \left\| \theta_P^* - \hat{\theta}_P \right\|_{\Sigma_P}^2 + 4 \left\| \hat{\theta}_P - \theta_P^* \right\|_{\Sigma_P}^2 \\ &\leq 16\epsilon_P, \end{aligned}$$

507 where at second and fourth step we evoke matrix concentration result from Lemma 3 of [5], at third
 508 step we use the fact that $\tilde{\theta}_{PQ}$ is the optimal solution within the set. Putting pieces together will
 509 conclude the proof. □

510 **Proof of Lemma 7:**

511 *Proof.* Since $\theta_P^* \notin \left\{ \theta : \left\| \theta - \hat{\theta}_Q \right\|_{\hat{\Sigma}_Q}^2 \leq 6\epsilon_Q \right\}$, then from Lemma 5 we know with probability at
 512 least $1 - 2\tau$, $\mathcal{E}_Q(\theta_P^*) \geq \epsilon_Q \geq \frac{1}{26} \mathcal{E}_Q(\theta)$ holds. □

513 **12 Proof of Convergence**

514 In this section we will present the proof of the convergence result. We first introduce some useful
 515 lemmas.

516 12.1 Technical Lemmas

517 The following proposition is standard and will be used to show the sub-gaussianity of the stochastic
518 gradients.

519 **Proposition 2.** *Given a random variable X , if $a \leq X \leq b$ with probability 1, then X is a $\frac{(b-a)^2}{4}$
520 sub-Gaussian random variable.*

521 The next lemma establishes the convergence of $\theta_{Q,t}$.

522 **Lemma 9** (High probability convergence of $\theta_{Q,t}$). *If we choose $\alpha_t = \frac{1}{\lambda_{\min}^+(\hat{\Sigma}_Q)(t+2\kappa_Q)}$, then with
523 probability at least $1 - \tau$, for any $t \geq 0$ we have:*

$$\hat{R}_Q(\theta_{Q,t}) - \hat{R}_Q(\hat{\theta}_Q) \leq \frac{\lambda_{\min}^+(\hat{\Sigma}_Q)(1+2\kappa_Q) \|\hat{\theta}_Q\|^2}{t+2\kappa_Q} + \frac{6\sigma_Q^2 \log(2/\tau)(\log t + 1)}{\lambda_{\min}^+(\hat{\Sigma}_Q)(t+2\kappa_Q)}$$

524 for $\sigma_Q^2 = \left(M_x^2 \left(\frac{1+\log(T+2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y \right) + M_x M_y \right)^2$.

525 *Proof.* We first examine the boundedness of $\|\theta_{Q,t}\|$. According to updating rule of $\theta_{Q,t}$ we have

$$\begin{aligned} \|\theta_{Q,t}\| &= \|\theta_{Q,t-1} - \alpha_t x_t (x_t^\top \theta_{Q,t-1} - y_t)\| \\ &\leq \|(\mathbf{I} - \alpha_t x_t x_t^\top) \theta_{Q,t-1}\| + \alpha_t \|x_t y_t\| \\ &\leq \underbrace{\|\theta_{Q,0}\|}_{=0} + \sum_{s=1}^t \frac{1}{\lambda_{\min}^+(\hat{\Sigma}_Q)(s+2\kappa_Q)} M_x \hat{M}_y \\ &\leq \frac{1 + \log(t+2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y. \end{aligned} \tag{8}$$

526 Hence we can compute the sub-Gaussian parameter. Notice that

$$\begin{aligned} \|\nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \hat{R}_Q(\theta_{Q,t})\| &\leq \|x_t x_t^\top - \hat{\Sigma}_Q\| \|\theta_{Q,t}\| + \left\| x_t y_t - \frac{1}{n_Q} \mathbf{X}_Q^\top \mathbf{y}_Q \right\| \\ &\leq 2M_x^2 \left(\frac{1 + \log(t+2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y \right) + 2M_x \hat{M}_y \\ &\leq 2M_x^2 \left(\frac{1 + \log(T+2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y \right) + 2M_x \hat{M}_y. \end{aligned}$$

527 According to Proposition 2, we know $\|\nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \hat{R}_Q(\theta_{Q,t})\|$ is

528 $\left(M_x^2 \left(\frac{1+\log(t+2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y \right) + M_x M_y \right)^2$ sub-Gaussian.

529 Now, we evoke the result from Theorem 3.7 from [27] that if the gradient noise is σ_Q sub-Gaussian,
530 then with our choice of α_t , with probability at least $1 - \tau$ it holds for any integer $t > 0$ that

$$\hat{R}_Q(\theta_{Q,t}) - \hat{R}_Q(\hat{\theta}_Q) \leq \frac{\lambda_{\min}^+(\hat{\Sigma}_Q)(1+2\kappa_Q) \|\hat{\theta}_Q\|^2}{t+2\kappa_Q} + \frac{6\sigma_Q^2 \log(2/\tau)(\log t + 1)}{\lambda_{\min}^+(\hat{\Sigma}_Q)(t+2\kappa_Q)}.$$

531 □

532 The next lemma proves the sub-Gaussianity of the stochastic gradient used to update θ_t .

Lemma 10. *Let*

$$g_\theta = \begin{cases} (1+\lambda) \nabla \ell(\theta; x, y), (x, y) \sim S_P, w.p. \frac{1}{1+\lambda} \\ (1+\lambda) \nabla \ell(\theta; x, y), (x, y) \sim S_Q, w.p. \frac{\lambda}{1+\lambda} \end{cases}$$

and

$$\delta = \left\| \nabla \hat{R}_P(\theta) + \lambda \nabla \hat{R}_Q(\theta) - g_\theta \right\|.$$

533 Then for any $\theta \in \left\{ \theta : \|\theta - \tilde{\theta}_{PQ}\|^2 \leq 2\rho^2 \right\}$ and $\lambda \in \left\{ \lambda : (\lambda - \lambda^*)^2 \leq 2\rho^2 \right\}$, we have
 534 $\mathbb{E}[\exp(\delta^2/\sigma_{PQ}^2)] \leq \exp(1)$ for $\sigma_{PQ}^2 = 256(1 + (\lambda^*)^2 + \rho^2) \hat{G}_\theta^2$.

535 *Proof.* We use $\xi = 1$ to denote the event that we sample from P , and otherwise from Q . For
 536 notational convenience we define $\delta_\mu \doteq \|\nabla \hat{R}_\mu(\theta) - \nabla \ell(\theta; x, y)\|$ where x, y is sampled from
 537 μ , for μ denoting either P or Q . We also define $\zeta_{PQ} \doteq \left\| \nabla \hat{R}_P(\theta) - \nabla \hat{R}_Q(\theta) \right\|$. Since
 538 $\delta = \left\| \nabla \hat{R}_P(\theta) + \lambda \nabla \hat{R}_Q(\theta) - (1 + \lambda)(\mathbb{I}\{\xi = 1\} \nabla \ell(\theta; x, y) + \mathbb{I}\{\xi = 0\} \nabla \ell(\theta; x, y)) \right\|$, we can
 539 verify that

$$\begin{aligned} & \mathbb{E} \exp(\delta^2/\sigma_{PQ}^2) \\ &= \frac{1}{1 + \lambda} \mathbb{E} \exp \left(\left\| \nabla \hat{R}_P(\theta) + \lambda \nabla \hat{R}_Q(\theta) - (1 + \lambda) \nabla \ell(\theta; x, y) \right\|^2 / \sigma_{PQ}^2 \right) \\ & \quad + \frac{\lambda}{1 + \lambda} \mathbb{E} \exp \left(\left\| \nabla \hat{R}_P(\theta) + \lambda \nabla \hat{R}_Q(\theta) - (1 + \lambda) \nabla \ell(\theta; x, y) \right\|^2 / \sigma_{PQ}^2 \right) \\ &\leq \frac{1}{1 + \lambda} \mathbb{E} \exp \left(\frac{2(1 + \lambda)^2 \delta_P^2}{\sigma_{PQ}^2} + \frac{2\lambda^2 \zeta_{PQ}^2}{\sigma_{PQ}^2} \right) + \frac{\lambda}{1 + \lambda} \mathbb{E} \exp \left(\frac{2\zeta_{PQ}^2}{\sigma_{PQ}^2} + \frac{2(1 + \lambda)^2 \delta_Q^2}{\sigma_{PQ}^2} \right) \\ &\leq \frac{1}{1 + \lambda} \left(\exp \left(\frac{2\lambda^2 \zeta_{PQ}^2}{\sigma_{PQ}^2} \right) \mathbb{E} \exp \left(2(1 + \lambda)^2 \frac{\delta_P^2}{\sigma_{PQ}^2} \right) + \lambda \exp \left(\frac{2\zeta_{PQ}^2}{\sigma_{PQ}^2} \right) \mathbb{E} \exp \left(\frac{2(1 + \lambda)^2 \delta_Q^2}{\sigma_{PQ}^2} \right) \right). \end{aligned}$$

540 Since $0 \leq \lambda \leq \sqrt{2\rho} + \lambda^*$, so we have

$$\begin{aligned} & \mathbb{E} \exp(\delta^2/\hat{\sigma}^2) \\ &\leq \frac{1}{1 + \lambda} \exp \left(2(2(\lambda^*)^2 + 4\rho^2) 4\hat{G}_\theta^2/\sigma_{PQ}^2 \right) \mathbb{E} \exp \left((4 + 4(2(\lambda^*)^2 + 4\rho^2)) \delta_P^2/\sigma_{PQ}^2 \right) \\ & \quad + \frac{\lambda_t}{1 + \lambda_t} \exp \left(4\hat{G}_\theta^2/\sigma_{PQ}^2 \right) \mathbb{E} \exp \left((4 + 4(2(\lambda^*)^2 + 4\rho^2)) \delta_Q^2/\sigma_{PQ}^2 \right). \end{aligned}$$

541 Due to our choice $\sigma_{PQ}^2 = 256(1 + (\lambda^*)^2 + \rho^2) \hat{G}_\theta^2$, we have

$$\mathbb{E} \exp(\delta^2/\sigma_{PQ}^2) \leq \frac{1}{1 + \lambda} \exp(1/8) (\exp(1))^{1/8} + \frac{\lambda}{1 + \lambda} \exp(1/8) (\exp(1))^{1/8} \leq \exp(1).$$

542 □

543 Then, we establish the convergence of the penalized objective, under the dynamic of Algorithm 1.

544 **Lemma 11.** For Algorithm 1, the following statement holds true for any $\lambda \geq 0$ with probability at
 545 least $1 - \tau$:

$$\begin{aligned} & \left(\hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\tilde{\theta}_{PQ}) \right) + \lambda (\hat{R}_Q(\bar{\theta}_T) - \hat{R}_Q(\hat{\theta}_Q) - 6\epsilon_Q) - \left(\frac{\gamma}{2} + \frac{1}{2\eta T} \right) \lambda^2 \\ &\leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \eta \hat{G}_\theta^2 + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} (\lambda^* + 2\sqrt{2}\rho) + \frac{\sqrt{2}\rho \sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\ & \quad + \frac{C_{PQ}(\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho)}{T} + \left(\frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} + \frac{C_{PQ}(\log(T + 2\kappa_Q) + 2)^2}{T} \right) \lambda, \end{aligned}$$

546 where $C_{PQ} := (1 + 2\kappa_Q) M_x^2 M_y^2 + \frac{6\sigma_Q^2 \log(2/\tau)}{\lambda_{\min}^+(\hat{\Sigma}_Q)}$, $\sigma_Q^2 = \left(M_x^2 \left(\frac{1 + \log(T + 2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y \right) + M_x M_y \right)^2$.

547 *Proof.* Define the constraint function $g(\theta) := \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) - 6\epsilon_Q$ and $L(\theta, \lambda) \doteq \hat{R}_P(\theta) +$
 548 $\lambda g(\theta) - \frac{\gamma\lambda^2}{2}$. We first show that $\|\theta_t - \tilde{\theta}_{PQ}\|^2 + \|\lambda_t - \lambda^*\|^2 \leq 2\rho^2$, for any $t \in [T]$. We prove this
 549 by induction. Assume this holding for t , and for $t+1$ we have

$$\begin{aligned}\|\theta_{t+1} - \tilde{\theta}_{PQ}\|^2 &= \|\theta_t - \eta g_\theta^t - \tilde{\theta}_{PQ}\|^2 \\ &= \|\theta_t - \tilde{\theta}_{PQ}\|^2 - 2\langle \eta g_\theta^t, \theta_t - \tilde{\theta}_{PQ} \rangle + \eta^2 \|g_\theta^t\|^2,\end{aligned}$$

550 where $g_\theta^t = \begin{cases} (1 + \lambda_t)\nabla\ell(\theta_t; x_t, y_t), (x_t, y_t) \sim S_P, w.p. \frac{1}{1+\lambda_t} \\ (1 + \lambda_t)\nabla\ell(\theta_t; x_t, y_t), (x_t, y_t) \sim S_Q, w.p. \frac{\lambda_t}{1+\lambda_t} \end{cases}.$

551 Then for $t+1$, we have:

$$\begin{aligned}\|\theta_{t+1} - \tilde{\theta}_{PQ}\|^2 &\leq \|\theta_t - \tilde{\theta}_{PQ}\|^2 - 2\eta_t \langle \nabla \hat{R}_P(\theta_t) + \lambda_t \nabla \hat{R}_Q(\theta_t), \theta_t - \tilde{\theta}_{PQ} \rangle \\ &\quad + 2\sqrt{2}\eta\delta_t\rho + \eta_t^2(1 + \lambda_t)^2\hat{G}_\theta^2 \\ &\leq \|\theta_t - \tilde{\theta}_{PQ}\|^2 - 2\eta \left(L(\theta_t, \lambda_t) - L(\tilde{\theta}_{PQ}, \lambda_t) \right) + 2\sqrt{2}\eta\delta_t\rho + \eta^2(1 + \lambda_t)^2\hat{G}_\theta^2\end{aligned}$$

where $\delta_t = \|\nabla \hat{R}_P(\theta_t) + \lambda_t \nabla \hat{R}_Q(\theta_t) - g_\theta^t\|$, and at last step we use the convexity of $L(\cdot, \lambda_t)$.
 According to Lemma 10 we know that

$$\mathbb{E}[g_\theta^t] = \nabla \hat{R}_P(\theta_t) + \lambda_t \nabla \hat{R}_Q(\theta_t), \mathbb{E}[\exp(\delta_t^2/\sigma_{PQ}^2)] \leq \exp(1)$$

552 . Similarly, we have:

$$\begin{aligned}|\lambda_{t+1} - \lambda^*|^2 &= |\lambda_t - \lambda^*|^2 + 2\eta \langle g_\lambda^t, \lambda_t - \lambda^* \rangle + \eta^2 |g_\lambda^t - \gamma\lambda_t|^2 \\ &\leq |\lambda_t - \lambda^*|^2 + 2\eta \langle \hat{R}_Q(\theta_t) - \hat{R}_Q(\hat{\theta}_Q) - \epsilon_Q - \gamma\lambda_t, \lambda_t - \lambda^* \rangle \\ &\quad + 2\sqrt{2}\eta r_t\rho + 2\sqrt{2}\eta h_t\rho + 2\eta^2\hat{G}_\lambda^2 \\ &\leq (1 - \gamma\eta)|\lambda_t - \lambda^*|^2 - 2\eta (L(\theta_t, \lambda^*) - L(\theta_t, \lambda_t)) \\ &\quad + 2\sqrt{2}\eta r_t\rho + 2\sqrt{2}\eta h_t\rho + 2\eta^2\hat{G}_\lambda^2,\end{aligned}$$

where

$$g_\lambda^t = \ell(\theta_t; x_t, y_t) - \ell(\theta_{Q,t}; x_t, y_t) - 6\epsilon_Q$$

and

$$r_t = \|\ell(\theta_t; x_t, y_t) - \ell(\theta_{Q,t}; x_t, y_t) - (\hat{R}_Q(\theta_t) - \hat{R}_Q(\theta_{Q,t}))\|, h_t = |\hat{R}_Q(\theta_{Q,t}) - \hat{R}_Q(\hat{\theta}_Q)|$$

and at last step we use the γ concavity of $L(\theta_t, \cdot)$. It is easy to see that

$$\mathbb{E}[g_\lambda^t] = \hat{R}_Q(\theta_t) - \hat{R}_Q(\hat{\theta}_Q) - \epsilon_Q, \mathbb{E}[\exp(r_t^2/\hat{G}_\lambda^2)] \leq \exp(1).$$

553 Putting pieces together we have

$$\begin{aligned}\|\theta_{t+1} - \tilde{\theta}_{PQ}\|^2 + |\lambda_{t+1} - \lambda^*|^2 &\leq \left(|\lambda_t - \lambda^*|^2 + \|\theta_t - \tilde{\theta}_{PQ}\|^2 \right) - 2\eta \left(L(\theta_t, \lambda^*) - L(\tilde{\theta}_{PQ}, \lambda_t) \right) \\ &\quad + 2\sqrt{2}\eta\delta_t\rho + 2\sqrt{2}\eta r_t\rho + 2\sqrt{2}\eta h_t\rho + 2\eta^2\hat{G}_\lambda^2 + \eta^2(1 + \lambda_t)^2\hat{G}_\theta^2 \\ &\leq \left(|\lambda_t - \lambda^*|^2 + \|\theta_t - \tilde{\theta}_{PQ}\|^2 \right) - \underbrace{2\eta \left(L(\theta_t, \lambda^*) - L(\tilde{\theta}_{PQ}, \lambda_t) \right)}_{\geq -\frac{\gamma(\lambda^*)^2}{2}} \\ &\quad + 2\eta^2\hat{G}_\lambda^2 + \eta^2(1 + \sqrt{2}\rho + \lambda^*)^2\hat{G}_\theta^2 + 2\sqrt{2}\eta\rho(\delta_t + r_t + h_t)\end{aligned}$$

554 where the last step is due to

$$L(\theta_t, \lambda^*) - L(\tilde{\theta}_{PQ}, \lambda_t) = \underbrace{\hat{R}_P(\theta_t) + \lambda^*g(\theta_t) - \left(\hat{R}_P(\tilde{\theta}_{PQ}) + \lambda_tg(\tilde{\theta}_{PQ}) \right)}_{\geq 0} - \frac{\gamma(\lambda^*)^2}{2} + \frac{\gamma\lambda_t^2}{2}.$$

555 Performing telescoping sum yields:

$$\begin{aligned} \left\| \theta_{t+1} - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_{t+1} - \lambda^*|^2 &\leq \left(\left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_0 - \lambda^*|^2 \right) + 2\eta^2 \hat{G}_\lambda^2 + \eta^2 (1 + \sqrt{2}\rho + \lambda^*)^2 \hat{G}_\theta^2 \\ &\quad + 2\sqrt{2}\eta\rho \sum_{s=0}^t \delta_s + 2\sqrt{2}\eta\rho \sum_{s=0}^t r_s + 2\sqrt{2}\eta\rho \sum_{s=0}^t h_s + T\gamma\eta(\lambda^*)^2. \end{aligned}$$

556 Due to Lemma 4 of [28], we know with probability $1 - \tau/2$,

$$\sum_{t=0}^{T-1} \delta_t \leq \sqrt{T\sigma_{PQ}^2} \sqrt{3 \log \frac{2}{\tau}}, \quad \sum_{t=0}^{T-1} r_t \leq \sqrt{T\hat{G}_\lambda^2} \sqrt{3 \log \frac{2}{\tau}}, \quad (9)$$

557 and also according to Lemma 9, $h_t \leq \frac{\lambda_{\min}^+(\hat{\Sigma}_Q)(1+2\kappa_Q)\|\hat{\theta}_Q\|^2}{t+2\kappa_Q} + \frac{6\sigma_Q^2 \log(2/\tau)(\log t+1)}{\lambda_{\min}^+(\hat{\Sigma}_Q)(t+2\kappa_Q)}$ for $\sigma_Q^2 =$
 558 $\left(M_x^2 \left(\frac{1+\log(T+2\kappa_Q)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} M_x M_y \right) + M_x M_y \right)^2$, which yields:

$$\begin{aligned} \sum_{s=0}^t h_s &= \sum_{s=0}^t \left(\frac{\lambda_{\min}^+(\hat{\Sigma}_Q)(1+2\kappa_Q)\|\hat{\theta}_Q\|^2}{s+2\kappa_Q} + \frac{6\sigma_Q^2 \log(2/\tau)(\log s+1)}{\lambda_{\min}^+(\hat{\Sigma}_Q)(s+2\kappa_Q)} \right) \\ &\leq \left(\lambda_{\min}^+(\hat{\Sigma}_Q)(1+2\kappa_Q)\|\hat{\theta}_Q\|^2 + \frac{6\sigma_Q^2 \log(2/\tau)(\log t+1)}{\lambda_{\min}^+(\hat{\Sigma}_Q)} \right) (\log(t+2\kappa_Q) + 2) \\ &\leq C_{PQ} \cdot (\log(t+2\kappa_Q) + 2)^2, \end{aligned} \quad (10)$$

559 where $C_{PQ} := (1+2\kappa_Q)M_x^2 M_y^2 + \frac{6\sigma_Q^2 \log(2/\tau)}{\lambda_{\min}^+(\hat{\Sigma}_Q)}$, and in the first inequality we use the fact $\sum_{s=1}^t \frac{1}{s} \leq$
 560 $1 + \int_1^t \frac{1}{s} \leq 1 + \log t$. Putting pieces together yields:

$$\begin{aligned} \left\| \theta_{t+1} - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_{t+1} - \lambda^*|^2 &\leq |\lambda_0 - \lambda^*|^2 + \left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2 + 2\eta^2 \hat{G}_\lambda^2 + \eta^2 (1 + \sqrt{2}\rho + \lambda^*)^2 \hat{G}_\theta^2 \\ &\quad + 4\sqrt{2}\eta\rho\sqrt{T}\sigma_Q \sqrt{3 \log \frac{2}{\tau}} + 2\sqrt{2}\eta\rho C_{PQ} (\log(t+2\kappa_Q) + 2)^2 + T\gamma\eta(\lambda^*)^2. \end{aligned} \quad (11)$$

561 Since we choose $\eta = \frac{c_\eta}{\sqrt{T}}$ and $\gamma = \hat{G}_\theta^2 \eta$, where

$$c_\eta \leq \min \left\{ \frac{\rho}{2\sqrt{2}\hat{G}_\lambda}, \frac{\rho}{2(1+\sqrt{2}\rho+\lambda^*)\hat{G}_\theta}, \frac{\rho}{16\sqrt{6}\sigma_{PQ}\sqrt{\log \frac{2}{\tau}}}, \frac{\rho}{4C_{PQ}} \right\},$$

562 we conclude that $\left\| \theta_{t+1} - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_{t+1} - \lambda^*|^2 \leq 2\rho^2$.

563 Now by similar analysis we have that for any $\lambda \geq 0$

$$\begin{aligned} \left\| \theta_{t+1} - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_{t+1} - \lambda|^2 &\leq \left\| \theta_t - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_t - \lambda|^2 - 2\eta(L(\theta_t, \lambda) - L(\tilde{\theta}_{PQ}, \lambda_t)) \\ &\quad + 2\eta^2 \hat{G}_\lambda^2 + \eta^2 (1 + \lambda_t)^2 \hat{G}_\theta^2 \\ &\quad + 2\eta \langle \ell(\theta_t; x_t, y_t) - \ell(\theta_{Q,t}; x_t, y_t) - (\ell(\theta_t) - \ell(\theta_{Q,t})), \lambda_t - \lambda \rangle \\ &\quad + 2\eta \langle \ell(\hat{\theta}_Q) - \ell(\theta_{Q,t}), \lambda_t - \lambda \rangle + 2\eta r_t \left\| \theta_t - \tilde{\theta}_{PQ} \right\| \\ &\leq \left\| \theta_t - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_t - \lambda|^2 - 2\eta(L(\theta_t, \lambda) - L(\tilde{\theta}_{PQ}, \lambda_t)) \\ &\quad + 2\eta^2 \hat{G}_\lambda^2 + \eta^2 (1 + \lambda_t)^2 \hat{G}_\theta^2 \\ &\quad + 2\eta r_t (\lambda_t + \lambda) + 2\eta h_t (\lambda_t + \lambda) + 2\sqrt{2}\eta\delta_t \rho. \end{aligned}$$

564 Since $|\lambda_t - \lambda^*| \leq \sqrt{2}\rho$ we know $\lambda_t \leq \lambda^* + \sqrt{2}\rho$. Hence we have

$$\begin{aligned} \left\| \theta_{t+1} - \tilde{\theta}_{PQ} \right\|^2 + |\lambda_{t+1} - \lambda|^2 &\leq |\lambda_t - \lambda|^2 + \left\| \theta_t - \tilde{\theta}_{PQ} \right\|^2 - 2\eta(L(\theta_t, \lambda) - L(\tilde{\theta}_{PQ}, \lambda_t)) \\ &\quad + 2\eta^2 \hat{G}_\lambda^2 + \eta^2(1 + \lambda_t)^2 \hat{G}_\theta^2 + 2\eta r_t (\lambda^* + \sqrt{2}\rho) + 2\eta h_t (\lambda^* + \sqrt{2}\rho) \\ &\quad + 2\eta r_t \lambda + 2\eta h_t \lambda + 2\sqrt{2}\eta \delta_t \rho. \end{aligned}$$

565 Performing telescoping sum yields:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} L(\theta_t, \lambda) - L(\tilde{\theta}_{PQ}, \lambda_t) &\leq \frac{1}{2\eta T} (|\lambda_0 - \lambda|^2 + \left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2) + \frac{1}{T} \eta \hat{G}_\lambda^2 + \frac{1}{2T} \eta \sum_{t=0}^{T-1} (1 + \lambda_t)^2 \hat{G}_\theta^2 \\ &\quad + \frac{1}{T} \sum_{t=0}^{T-1} r_t (\lambda^* + \sqrt{2}\rho) + \frac{1}{T} \sum_{t=0}^{T-1} h_t (\lambda^* + \sqrt{2}\rho) + \frac{1}{T} \sum_{t=0}^{T-1} r_t \lambda \\ &\quad + \frac{1}{T} \sum_{t=0}^{T-1} h_t \lambda + \sqrt{2}\rho \frac{1}{T} \sum_{t=0}^{T-1} \delta_t. \end{aligned}$$

566 By the definition of Lagrangian, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} (\hat{R}_P(\theta_t) + \lambda g(\theta_t) - \frac{\gamma}{2} \lambda^2 - \hat{R}_P(\tilde{\theta}_{PQ}) - \underbrace{\lambda_t g(\tilde{\theta}_{PQ})}_{\leq 0} + \frac{\gamma}{2} \lambda_t^2) \\ \leq \frac{1}{2\eta T} (|\lambda_0 - \lambda|^2 + \left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2) + \frac{1}{T} \eta \hat{G}_\lambda^2 + \frac{1}{2T} \eta \sum_{t=0}^{T-1} (1 + \lambda_t)^2 \hat{G}_\theta^2 \\ + \frac{1}{T} \sum_{t=0}^{T-1} r_t (\lambda^* + \sqrt{2}\rho) + \frac{1}{T} \sum_{t=0}^{T-1} h_t (\lambda^* + \sqrt{2}\rho) + \frac{1}{T} \sum_{t=0}^{T-1} r_t \lambda + \frac{1}{T} \sum_{t=0}^{T-1} h_t \lambda + \sqrt{2}\rho \frac{1}{T} \sum_{t=0}^{T-1} \delta_t. \end{aligned}$$

567 Evoking the bound from (9) and (10) yields:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} (\hat{R}_P(\theta_t) + \lambda g(\theta_t) - \frac{\gamma}{2} \lambda^2 - \hat{R}_P(\tilde{\theta}_{PQ}) + \frac{\gamma}{2} \lambda_t^2) \\ \leq \frac{1}{2\eta T} (|\lambda_0 - \lambda|^2 + \left\| \theta_0 - \tilde{\theta}_{PQ} \right\|^2) + \frac{1}{T} \eta \hat{G}_\lambda^2 + \frac{1}{2T} \eta \sum_{t=0}^{T-1} (1 + \lambda_t)^2 \hat{G}_\theta^2 \\ + \frac{1}{T} \sqrt{T \hat{G}_\lambda^2} \sqrt{3 \log \frac{2}{\tau}} (\lambda^* + 2\sqrt{2}\rho) + \frac{1}{T} C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho) \\ + \frac{1}{T} \sqrt{T \hat{G}_\lambda^2} \sqrt{3 \log \frac{2}{\tau}} \lambda + \frac{1}{T} C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 \lambda + \frac{\sqrt{2}\rho \sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}}. \end{aligned}$$

568 Plugging in $\lambda_0 = 0, \theta_0 = \mathbf{0}$ and re-arranging the terms yields:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} (\hat{R}_P(\theta_t) - \hat{R}_P(\tilde{\theta}_{PQ})) + \frac{1}{T} \sum_{t=0}^{T-1} \lambda g(\theta_t) - \left(\frac{\gamma}{2} + \frac{1}{2\eta T} \right) \lambda^2 \\ \leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \frac{1}{2T} \sum_{t=0}^{T-1} (\eta(1 + \lambda_t)^2 \hat{G}_\theta^2 - \gamma \lambda_t^2) + \frac{\sqrt{2}\rho \sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\ + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} (\lambda^* + 2\sqrt{2}\rho) + \frac{1}{T} C_{PQ} (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + 2\sqrt{2}\rho) \\ + \left(\frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} + \frac{C_{PQ} (\log(T + 2\kappa_Q) + 2)^2}{T} \right) \lambda. \end{aligned}$$

569 By our choice, $\gamma = \hat{G}_\theta^2 \eta$, so we have

$$\begin{aligned} & \frac{1}{T} \sum_{t=0}^{T-1} \left(\hat{R}_P(\theta_t) - \hat{R}_P(\tilde{\theta}_{PQ}) \right) + \frac{1}{T} \sum_{t=0}^{T-1} \lambda g(\theta_t) - \left(\frac{\gamma}{2} + \frac{1}{2\eta T} \right) \lambda^2 \\ & \leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \eta \hat{G}_\theta^2 + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} (\lambda^* + \sqrt{2}\rho) + \frac{\sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\ & + \frac{1}{T} C_{PQ} (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho) + \left(\frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} + \frac{C_{PQ} (\log(T + 2\kappa_Q) + 2)^2}{T} \right) \lambda. \end{aligned}$$

570 Define $\hat{\theta}_T = \frac{1}{T} \sum_{t=0}^{T-1} \theta_t$, and then by Jensen's inequality we have

$$\begin{aligned} & \left(\hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\tilde{\theta}_{PQ}) \right) + \lambda (\hat{R}_Q(\bar{\theta}_T) - \hat{R}_Q(\hat{\theta}_Q) - 6\epsilon_Q) - \left(\frac{\gamma}{2} + \frac{1}{2\eta T} \right) \lambda^2 \\ & \leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \eta \hat{G}_\theta^2 + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} (\lambda^* + \sqrt{2}\rho) + \frac{\sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\ & + \frac{1}{T} C_{PQ} (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho) + \left(\frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} + \frac{C_{PQ} (\log(T + 2\kappa_Q) + 2)^2}{T} \right) \lambda. \end{aligned}$$

571

□

572 12.2 Proof of Theorem 2

573 *Proof.* Note that Lemma 11 holds for any $\lambda \geq 0$. Now let's discuss by cases. If $\bar{\theta}_T$ is in the constraint
574 set, then $\hat{\theta}_{PQ} = \bar{\theta}_T$ and we simply set $\lambda = 0$ and get the convergence:

$$\begin{aligned} & \hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \\ & \leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \eta \hat{G}_\theta^2 + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} (\lambda^* + 2\sqrt{2}\rho) + \frac{\sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\ & + \frac{1}{T} C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho). \end{aligned}$$

575 If $\bar{\theta}_T$ is not in the constraint set, we set $\lambda = \frac{\hat{R}_Q(\bar{\theta}_T) - \hat{R}_Q(\hat{\theta}_Q) - 6\epsilon_Q}{\gamma + \frac{1}{\eta T}}$, and define $g(\theta) := \hat{R}_Q(\theta) -$

576 $\hat{R}_Q(\hat{\theta}_Q) - 6\epsilon_Q$ for notational simplicity, which yields:

$$\begin{aligned} & (\hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\tilde{\theta}_{PQ})) + \frac{(g(\bar{\theta}_T))^2}{2(\gamma + \frac{1}{\eta T})} \\ & \leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \eta \hat{G}_\theta^2 + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} (\lambda^* + 2\sqrt{2}\rho) + \frac{\sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\ & + \frac{1}{T} C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho) \\ & + \left(\frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} + \frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2}{T} \right) \left| \frac{g(\bar{\theta}_T)}{\gamma + \frac{1}{\eta T}} \right|. \end{aligned} \quad (12)$$

577 Since $\bar{\theta}_T$ is not in the constraint set and $\hat{\theta}_{PQ}$ is the projection of it onto inexact constraint set
578 $\tilde{g}(\theta) := \hat{R}_Q(\theta) - \hat{R}_Q(\theta_{Q,T}) - 3\epsilon_Q \leq 0$, by KKT condition we know $\tilde{g}(\hat{\theta}_{PQ}) = 0$ and $\bar{\theta}_T - \hat{\theta}_{PQ} =$
579 $s \cdot \nabla \tilde{g}(\hat{\theta}_{PQ})$ for some $s > 0$. Defining $\Delta := 3\epsilon_Q - (\hat{R}_Q(\theta_{Q,T}) - \hat{R}_Q(\hat{\theta}_Q))$, and due to our choice

580 of T we know $\Delta \geq 0$. Then we have

$$\begin{aligned}
g(\bar{\theta}_T) &= g(\bar{\theta}_T) - \tilde{g}(\hat{\theta}_{PQ}) \\
&= \tilde{g}(\bar{\theta}_T) - \tilde{g}(\hat{\theta}_{PQ}) - (\tilde{g}(\bar{\theta}_T) - g(\bar{\theta}_T)) \\
&= \tilde{g}(\bar{\theta}_T) - \tilde{g}(\hat{\theta}_{PQ}) - (\hat{R}_Q(\hat{\theta}_Q) - \hat{R}_Q(\theta_{Q,T}) + 3\epsilon_Q) \\
&\geq \langle \nabla \tilde{g}(\hat{\theta}_{PQ}), \bar{\theta}_T - \hat{\theta}_{PQ} \rangle - \Delta = \left\| \nabla \tilde{g}(\hat{\theta}_{PQ}) \right\| \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| - \Delta
\end{aligned}$$

581 where the inequality is due to convexity of $g(\cdot)$. Evoking Lemma 8 with $\epsilon = 3\epsilon_Q + \hat{R}_Q(\theta_{Q,T}) -$
582 $\hat{R}_Q(\hat{\theta}_Q)$ gives that $\min_{\tilde{g}(\theta)=0} \left\| \nabla \tilde{g}(\hat{\theta}_{PQ}) \right\| \geq \sqrt{\lambda_{\min}^+(\hat{\Sigma}_Q)(3\epsilon_Q + \hat{R}_Q(\theta_{Q,T}) - \hat{R}_Q(\hat{\theta}_Q))} \geq$
583 $\sqrt{\lambda_{\min}^+(\hat{\Sigma}_Q)3\epsilon_Q}$, so $g(\bar{\theta}_T) \geq \sqrt{\lambda_{\min}^+(\hat{\Sigma}_Q)3\epsilon_Q} \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| - \Delta$.

584 On the other hand, since $\hat{\theta}_{PQ}$ is the projection of $\bar{\theta}_T$ onto constraint set, and $\tilde{\theta}_{PQ}$ is in the constraint
585 set, we know

$$\left\| \hat{\theta}_{PQ} - \tilde{\theta}_{PQ} \right\|^2 \leq \left\| \bar{\theta}_T - \tilde{\theta}_{PQ} \right\|^2 \leq 2\rho^2.$$

586 Hence $\hat{\theta}_{PQ}$ also falls in the set $\left\{ \theta : \left\| \theta - \tilde{\theta}_{PQ} \right\|^2 \leq 2\rho^2 \right\}$, so the gradient upper bound \hat{G}_θ applies
587 to $\hat{\theta}_{PQ}$. Hence we also know

$$\begin{aligned}
g(\bar{\theta}_T) &= g(\bar{\theta}_T) - \tilde{g}(\hat{\theta}_{PQ}) \\
&= \hat{R}_Q(\bar{\theta}_T) - \hat{R}_Q(\hat{\theta}_Q) - 6\epsilon_Q - (\hat{R}_Q(\hat{\theta}_{PQ}) - \hat{R}_Q(\theta_{Q,T}) - 3\epsilon_Q) \\
&\leq \hat{G}_\theta \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| + \underbrace{\hat{R}_Q(\theta_{Q,T}) - \hat{R}_Q(\hat{\theta}_Q) - 3\epsilon_Q}_{\leq 0}.
\end{aligned}$$

588 Plugging the upper and lower bound of $g(\bar{\theta}_T)$ into (12) yields:

$$\begin{aligned}
&(\hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\tilde{\theta}_{PQ})) + \sqrt{T} \frac{\lambda_{\min}^+(\hat{\Sigma}_Q)3\epsilon_Q \left(\left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| - \Delta \right)^2}{4(c_\eta \hat{G}_\theta^2 + \frac{1}{c_\eta})} \\
&\leq \frac{\rho^2}{\eta T} + \eta \hat{G}_\lambda^2 + \eta \hat{G}_\theta^2 + \frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \left(\lambda^* + 2\sqrt{2}\rho \right) + \frac{\sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \\
&\quad + \frac{1}{T} C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 \left(\lambda^* + \sqrt{2}\rho \right) \\
&\quad + \left(\frac{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} + \frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2}{T} \right) \frac{\sqrt{T}}{c_\eta \hat{G}_\theta^2 + \frac{1}{c_\eta}} \hat{G}_\theta \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\|.
\end{aligned}$$

589 Notice the following decomposition:

$$\hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\tilde{\theta}_{PQ}) \geq \hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\hat{\theta}_{PQ}) \geq -\hat{G}_\theta \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\|.$$

590 Also notice the fact $(a - b)^2 \geq \frac{1}{2}a^2 - b^2$ holding for any $a > 0, b > 0$, we know

$$\left(\left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| - \Delta \right)^2 \geq \frac{1}{2} \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\|^2 - \Delta^2.$$

591 Putting pieces together yield the following inequality:

$$a \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\|^2 - b \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| - c \leq 0,$$

where:

$$\begin{aligned} a &= \frac{3\sqrt{T}\lambda_{\min}^+(\hat{\Sigma}_Q)\epsilon_Q}{8(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})} \\ b &= \frac{\hat{G}_\theta}{(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})} \left(\hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}} + \frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2}{\sqrt{T}} \right) + \hat{G}_\theta, \\ c &= \frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}} (\lambda^* + 2\sqrt{2}\rho) + \sqrt{2}\rho\sigma_{PQ} \sqrt{3\log \frac{2}{\tau}}}{\sqrt{T}} \\ &\quad + \frac{C \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho)}{T} + \sqrt{T} \frac{3\lambda_{\min}^+(\hat{\Sigma}_Q)\epsilon_Q \Delta^2}{4(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})}. \end{aligned}$$

592 Assume $T \geq \left(\frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2}{\hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}}} \right)^2$, so $b \leq \frac{2\hat{G}_\theta \hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}}}{(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})} + \hat{G}_\theta$. Solving the above quadratic
593 inequality yields:

$$\begin{aligned} \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| &\leq \frac{b + \sqrt{b^2 + 4ac}}{2a} \leq \frac{b}{a} + \sqrt{\frac{c}{a}} \\ &\leq \frac{8(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})\hat{G}_\theta + 16\hat{G}_\theta\hat{G}_\lambda\sqrt{3\log \frac{2}{\tau}}}{3\lambda_{\min}^+(\hat{\Sigma}_Q)\epsilon_Q\sqrt{T}} + \frac{1}{2}\Delta \\ &\quad + \sqrt{\frac{8(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})}{3\sqrt{T}\lambda_{\min}^+(\hat{\Sigma}_Q)\epsilon_Q} \sqrt{\frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}} (\lambda^* + 2\sqrt{2}\rho) + \sqrt{2}\rho\sigma_{PQ} \sqrt{3\log \frac{2}{\tau}}}{\sqrt{T}}}} \\ &\quad + \sqrt{\frac{8(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})}{3\sqrt{T}\lambda_{\min}^+(\hat{\Sigma}_Q)\epsilon_Q} \sqrt{\frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho)}{T}}} \\ &= \frac{16(c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta})\hat{G}_\theta + 16\hat{G}_\theta\hat{G}_\lambda\sqrt{3\log \frac{2}{\tau}}}{\lambda_{\min}^+(\hat{\Sigma}_Q)\epsilon_Q\sqrt{T}} + \frac{1}{2}\Delta + \frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho)}{2T} \\ &\quad + \frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}} (\lambda^* + 2\sqrt{2}\rho) + \sqrt{2}\rho\sigma_{PQ} \sqrt{3\log \frac{2}{\tau}}}{2\sqrt{T}} \end{aligned}$$

594 where at the last step we use the fact $\sqrt{ab} \leq \frac{a^2+b^2}{2}$. Finally, note the following decomposition:

$$\begin{aligned} \hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\bar{\theta}_{PQ}) &= \hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\bar{\theta}_T) + \hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\bar{\theta}_{PQ}) \\ &\leq \hat{G}_\theta \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| + \hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\bar{\theta}_{PQ}) \\ &\leq \left(\hat{G}_\theta + \frac{2\hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}}}{c_\eta\hat{G}_\theta^2 + \frac{1}{c_\eta}} \hat{G}_\theta \right) \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| \\ &\quad + \frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \hat{G}_\lambda \sqrt{3\log \frac{2}{\tau}} (\lambda^* + 2\sqrt{2}\rho) + \sqrt{2}\rho\sigma_{PQ} \sqrt{3\log \frac{2}{\tau}}}{\sqrt{T}} \\ &\quad + \frac{C_{PQ} \cdot (\log(T + 2\kappa_Q) + 2)^2 (\lambda^* + \sqrt{2}\rho)}{T}. \end{aligned}$$

595 Since we assume $T \geq \left(\frac{C_{PQ} \cdot (\log(T+2\kappa_Q))^2}{\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}} \right)^2$, we know

$$\begin{aligned} \hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) &\leq \hat{G}_\theta \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| + \hat{R}_P(\bar{\theta}_T) - \hat{R}_P(\tilde{\theta}_{PQ}) \\ &\leq \left(\hat{G}_\theta + \frac{2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{c_\eta \hat{G}_\theta^2 + \frac{1}{c_\eta}} \hat{G}_\theta \right) \left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\| \\ &\quad + \frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + 2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}} (\lambda^* + 2\sqrt{2}\rho) + \sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}}. \end{aligned}$$

596 Plugging bound of $\left\| \bar{\theta}_T - \hat{\theta}_{PQ} \right\|$ and Lemma 9 yields:

$$\begin{aligned} &\hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \\ &\leq \left(\hat{G}_\theta + \frac{2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{c_\eta \hat{G}_\theta^2 + \frac{1}{c_\eta}} \hat{G}_\theta \right) \left(\frac{16(c_\eta \hat{G}_\theta^2 + \frac{1}{c_\eta}) \hat{G}_\theta + 16\hat{G}_\theta \hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\lambda_{\min}^+(\hat{\Sigma}_Q) \epsilon_Q \sqrt{T}} + \frac{2C_{PQ} \log T}{(T + 2\kappa_Q)} \right) \\ &\quad + 48 \left(\hat{G}_\theta + \frac{\hat{G}_\lambda \sqrt{\log \frac{2}{\tau}}}{c_\eta \hat{G}_\theta^2 + \frac{1}{c_\eta}} \hat{G}_\theta \right) \left(\frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \hat{G}_\lambda \sqrt{\log \frac{2}{\tau}} (\lambda^* + \rho) + \rho\sigma_{PQ} \sqrt{\log \frac{2}{\tau}}}{\sqrt{T}} \right). \end{aligned}$$

597 Now we simplify the above bound. By the definition of c_η we know $c_\eta \leq \frac{1}{\hat{G}_\theta}$, so we have

$$\begin{aligned} &\hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \\ &\leq \left(\hat{G}_\theta + \frac{2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{c_\eta \hat{G}_\theta + 1} \right) \left(\frac{16(\hat{G}_\theta + \frac{1}{c_\eta}) \hat{G}_\theta + 16\hat{G}_\theta \hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\lambda_{\min}^+(\hat{\Sigma}_Q) \epsilon_Q \sqrt{T}} + \frac{2C_{PQ} \log T}{(T + 2\kappa_Q)} \right) \\ &\quad + 48 \left(\hat{G}_\theta + \frac{\hat{G}_\lambda \sqrt{\log \frac{2}{\tau}}}{c_\eta \hat{G}_\theta + 1} \right) \left(\frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \hat{G}_\lambda \sqrt{\log \frac{2}{\tau}} (\lambda^* + \rho) + \rho\sigma_{PQ} \sqrt{\log \frac{2}{\tau}}}{\sqrt{T}} \right) \\ &\leq \left(\hat{G}_\theta + 2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}} \right) \left(\frac{16(\hat{G}_\theta + \frac{1}{c_\eta}) \hat{G}_\theta + 16\hat{G}_\theta \hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\lambda_{\min}^+(\hat{\Sigma}_Q) \epsilon_Q \sqrt{T}} + \frac{2C_{PQ} \log T}{(T + 2\kappa_Q)} \right) \\ &\quad + 48 \left(\hat{G}_\theta + \hat{G}_\lambda \sqrt{\log \frac{2}{\tau}} \right) \left(\frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + ((\lambda^* + \rho) \hat{G}_\lambda + \rho\sigma_{PQ}) \sqrt{\log \frac{2}{\tau}}}{\sqrt{T}} \right). \end{aligned}$$

598 Again recall we choose: $T \geq \left(\frac{C_{PQ} \cdot (\log(T+2\kappa_Q))^2}{\hat{G}_\lambda \sqrt{\log 1/\tau}} \right)^2$, so $\frac{2C_{PQ} \log T}{(T+2\kappa_Q)} \leq \frac{\hat{G}_\lambda \sqrt{\log 1/\tau}}{\sqrt{T}}$. So we have

$$\begin{aligned} &\hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \\ &\leq \left(\hat{G}_\theta + 2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}} \right) \left(\frac{32\frac{\hat{G}_\theta}{c_\eta} + 16\hat{G}_\theta \hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}}}{\lambda_{\min}^+(\hat{\Sigma}_Q) \epsilon_Q \sqrt{T}} + \frac{\hat{G}_\lambda \sqrt{\log \frac{1}{\tau}}}{\sqrt{T}} \right) \\ &\quad + 2 \left(\hat{G}_\theta + 2\hat{G}_\lambda \sqrt{3 \log \frac{2}{\tau}} \right) \left(\frac{\frac{\rho^2}{c_\eta} + c_\eta(\hat{G}_\lambda^2 + \hat{G}_\theta^2) + \sqrt{2}\rho\sigma_{PQ} \sqrt{3 \log \frac{2}{\tau}}}{\sqrt{T}} \right) \\ &\lesssim \left(\hat{G}_\theta + \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}} \right) \left(\frac{\frac{\hat{G}_\theta}{c_\eta} + \hat{G}_\theta \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}}}{\lambda_{\min}^+(\hat{\Sigma}_Q) \epsilon_Q \sqrt{T}} + \frac{((\lambda^* + \rho) \hat{G}_\lambda + \rho\sigma_{PQ}) \sqrt{\log \frac{1}{\tau}} + \frac{\rho^2}{c_\eta} + c_\eta \hat{G}_\lambda^2}{\sqrt{T}} \right). \end{aligned}$$

Finally, by definition of c_η we know $\frac{\rho^2}{c_\eta} \geq c_\eta \hat{G}_\lambda^2$, $\frac{\rho^2}{c_\eta} \geq \rho \hat{G}_\lambda$ and $\frac{\rho^2}{c_\eta} \geq \rho 16\sqrt{6}\sigma_{PQ}\sqrt{\log \frac{2}{\tau}}$ which concludes the proof:

$$\hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \lesssim \left(\hat{G}_\theta + \hat{G}_\lambda \sqrt{\log \frac{1}{\tau}} \right) \left(\frac{\hat{G}_\theta}{\lambda_{\min}^+(\hat{\Sigma}_Q) \epsilon_Q \sqrt{T}} + \frac{\lambda^* \hat{G}_\lambda \sqrt{\log \frac{1}{\tau} + \frac{\rho^2}{c_\eta}}}{\sqrt{T}} \right).$$

□

13 Proof of the Results of General Loss

In this section we provide the missing proofs in Section 8. We first introduce the following lemma which establishes the convergence of auxiliary iterates $\theta_{Q,t}$ to Q ERM model.

Lemma 12 (High probability convergence of $\theta_{Q,t}$). *If we choose $\alpha_t = \frac{1}{m_1} \cdot \frac{1}{t+2\kappa}$, then with probability at least $1 - \tau$, for any $t \geq 0$ we have:*

$$\hat{R}_Q(\theta_{Q,t}) - \hat{R}_Q(\hat{\theta}_Q) \leq \frac{m_1(1+2\kappa_Q)(\frac{2}{m_1^2}\|\nabla \hat{R}_Q(\theta_{Q,0})\|^2 + 2\|\theta_{Q,0}\|^2)}{t+2\kappa} + \frac{6\hat{G}_\theta^2 \log(2/\tau)(\log t + 1)}{m_1(t+2\kappa)}.$$

Proof. We first examine the boundedness of $\|\theta_{Q,t}\|$. Define $e_t \doteq \nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \ell(\theta_{Q,0}; x_t, y_t)$. According to updating rule of $\theta_{Q,t}$ we have

$$\begin{aligned} \|\theta_{Q,t+1} - \theta_{Q,0}\| &\leq \|\theta_{Q,t} - \alpha_t \nabla \ell(\theta_{Q,t}; x_t, y_t) - (\theta_{Q,0} - \alpha_t \nabla \ell(\theta_{Q,0}; x_t, y_t))\| \\ &\quad + \alpha_t \|\nabla \ell(\theta_{Q,0}; x_t, y_t)\| \\ &= \sqrt{\|\theta_{Q,t} - \theta_{Q,0}\|^2 - 2\alpha_t \langle e_t, \theta_{Q,t} - \theta_{Q,0} \rangle + \alpha_t^2 \|e_t\|^2} \\ &\quad + \alpha_t \|\nabla \ell(\theta_{Q,0}; x_t, y_t)\| \\ &\leq \|\theta_{Q,t} - \theta_{Q,0}\| + \alpha_t \|\nabla \ell(\theta_{Q,0}; x_t, y_t)\| \\ &\leq \sum_{j=0}^t \alpha_j^2 \|\nabla \ell(\theta_{Q,0}; x_j, y_j)\|^2 \\ &\leq \sum_{j=0}^t \frac{1}{m_1} \cdot \frac{1}{t+2\kappa} \sup_{(x,y) \in S_Q} \|\nabla \ell(\theta_{Q,0}; x, y)\| \\ &\leq \frac{1 + \log(T+2\kappa)}{m_1} \sup_{(x,y) \in S_Q} \|\nabla \ell(\theta_{Q,0}; x, y)\| \end{aligned} \tag{13}$$

where the third step is due to the co-coercivity of the gradient of the convex and smooth functions:

$$\langle \nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \ell(\theta_{Q,0}; x_t, y_t), \theta_{Q,t} - \theta_{Q,0} \rangle \geq \frac{1}{m_2} \|\nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \ell(\theta_{Q,0}; x_t, y_t)\|^2.$$

Hence we can bound $\|\theta_{Q,t}\|$ as

$$\begin{aligned} \|\theta_{Q,t}\| &\leq \|\theta_{Q,t} - \theta_{Q,0}\| + \|\theta_{Q,0}\| \\ &\leq \frac{1 + \log(T+2\kappa)}{m_1} \sup_{(x,y) \in S_Q} \|\nabla \ell(\theta_{Q,0}; x, y)\|. \end{aligned}$$

Hence we can compute sub-Gaussian parameter. By the definition of \hat{G}_θ

$$\left\| \nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \hat{R}_Q(\theta_{Q,t}) \right\| \leq 2 \sup_{(x,y) \in S_Q} \|\nabla \ell(\theta_{Q,t}; x, y)\| \leq 2\hat{G}_\theta$$

According to Proposition 2, we know $\left\| \nabla \ell(\theta_{Q,t}; x_t, y_t) - \nabla \hat{R}_Q(\theta_{Q,t}) \right\|$ is \hat{G}_θ^2 sub-Gaussian.

Now, we evoke the result from Theorem 3.7 from [27] that if the gradient noise is \hat{G}_θ^2 sub-Gaussian, then with our choice of α_t , with probability at least $1 - \tau$ it holds for any integer $t > 0$ that

$$\hat{R}_Q(\theta_{Q,t}) - \hat{R}_Q(\hat{\theta}_Q) \leq \frac{m_1(1 + 2\kappa_Q)\|\hat{\theta}_Q\|^2}{t + 2\kappa} + \frac{6\hat{G}_\theta^2 \log(2/\tau)(\log t + 1)}{m_1(t + 2\kappa)}.$$

We further bound $\|\hat{\theta}_Q\|^2$ as

$$\begin{aligned} \|\hat{\theta}_Q\|^2 &\leq 2\|\hat{\theta}_Q - \theta_{Q,0}\|^2 + 2\|\theta_{Q,0}\|^2 \\ &\leq \frac{4}{m_1}(\hat{R}_Q(\theta_{Q,0}) - \hat{R}_Q(\hat{\theta}_Q)) + 2\|\theta_{Q,0}\|^2 \\ &\leq \frac{4}{m_1}(\hat{R}_Q(\theta_{Q,0}) - \min_{\theta \in \mathbb{R}^D} \hat{R}_Q(\theta)) + 2\|\theta_{Q,0}\|^2 \\ &\leq \frac{4}{m_1} \cdot \frac{1}{2m_1} \left\| \nabla \hat{R}_Q(\theta_{Q,0}) \right\|^2 + 2\|\theta_{Q,0}\|^2 \end{aligned}$$

which concludes the proof. \square

13.1 Proof of Theorem 4

Proof. The proof is almost identical to that of Theorem 2. In Section 12.2, choosing $C_{PQ} = m_1(1 + 2\kappa)(\frac{2}{m_1^2}\|\nabla \hat{R}_Q(\theta_{Q,0})\|^2 + 2\|\theta_{Q,0}\|^2) + \frac{6\hat{G}_\theta^2 \log(2/\tau)}{m_1}$ and plugging in $\tilde{g}(\theta) = \hat{R}_Q(\theta) - \hat{R}_Q(\theta_{Q,T}) - 2\epsilon_Q$ will conclude the proof. \square

13.2 Proof of Proposition 1

Proof. By the standard Rademacher complexity analysis (see [25]) we know:

$$\sup_{h \in \mathcal{H}} |R_\mu(\theta) - \hat{R}_\mu(\theta)| \leq 2\mathcal{R}_n(\ell \circ \mathcal{H}) + M_\ell \sqrt{\frac{\log \frac{2}{\tau}}{2n_\mu}}.$$

Plugging in the upper bound of $\mathcal{R}_n(\ell \circ \mathcal{H})$ from Assumption 5 concludes the proof. \square

13.3 Proof of Corollary 1

Proof. First, since $\hat{\theta}_{PQ} \in \{\theta : \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) \leq 3\epsilon_Q\}$, then by Proposition 1 and our choice of ϵ_Q we know $\mathcal{E}_Q(\hat{\theta}_{PQ}) \leq 5\epsilon_Q$ with probability at least $1 - \tau$ over the randomness of S_Q .

Then we discuss by cases. If $\theta_P^* \in \{\theta : \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) \leq 3\epsilon_Q\}$, then since $\hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) \leq \epsilon_P$, by Proposition 1 and our choice of ϵ_P we know with probability at least $1 - 2\tau$ over the randomness of S_P and Algorithm 3,

$$\begin{aligned} \mathcal{E}_P(\hat{\theta}_{PQ}) &= R_P(\hat{\theta}_{PQ}) - R_P(\theta_P^*) \\ &= R_P(\hat{\theta}_{PQ}) - \hat{R}_P(\hat{\theta}_{PQ}) + \hat{R}_P(\hat{\theta}_{PQ}) - \hat{R}_P(\tilde{\theta}_{PQ}) + \underbrace{\hat{R}_P(\tilde{\theta}_{PQ}) - \hat{R}_P(\theta_P^*)}_{\leq 0} \\ &\quad + \hat{R}_P(\theta_P^*) - R_P(\theta_P^*) \leq 3\epsilon_P. \end{aligned}$$

Hence $\mathcal{E}_Q(\hat{\theta}_{PQ}) \leq \delta(3\epsilon_P)$.

If $\theta_P^* \notin \{\theta : \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) \leq 3\epsilon_Q\}$, then we know $\mathcal{E}_Q(\theta_P^*) \geq \epsilon_Q$ with probability at least $1 - \tau$ over the randomness of S_Q . This is because for any θ such that $\mathcal{E}_Q(\theta) \leq \epsilon_Q$, it must be in the set $\{\theta : \hat{R}_Q(\theta) - \hat{R}_Q(\hat{\theta}_Q) \leq 3\epsilon_Q\}$. To see this, note that

$$\hat{\mathcal{E}}_Q(\theta) \leq \mathcal{E}_Q(\theta) + 2\epsilon_Q \leq 3\epsilon_Q.$$

636 Hence we know

$$\mathcal{E}_Q(\hat{\theta}_{PQ}) \leq 5\epsilon_Q \leq 5\mathcal{E}_Q(\theta_P^*) \leq 5\delta(\epsilon_P).$$

637 Putting piece together we have with probability at least $1 - 3\tau$, it holds that

$$\mathcal{E}_Q(\hat{\theta}_{PQ}) \leq 5 \min \{ \epsilon_Q, \delta(3\epsilon_P) \}.$$

638

□

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We clearly claim our contributions and scope in both abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We point out that our theoretical results are for convex functions which can be a potential limitation.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: In our Theorems, we clearly list the assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We report clear experiments setup in the Experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not make our code public at this moment. The data used in the experiments are open source data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We discuss this in the Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report them in the Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report the compute resources in the Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This paper conforms, in every respect, with the NeurIPS Code of Ethic.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: This is a theoretical paper so we are not aware of any societal impacts of it.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: This is a theoretical paper so we are not aware of any risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the references for the datasets used in the experiments.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not have crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

948 Question: Does the paper describe the usage of LLMs if it is an important, original, or
949 non-standard component of the core methods in this research? Note that if the LLM is used
950 only for writing, editing, or formatting purposes and does not impact the core methodology,
951 scientific rigorousness, or originality of the research, declaration is not required.

952 Answer: [NA]

953 Justification: We do not use LLM.

954 Guidelines:

- 955 • The answer NA means that the core method development in this research does not
956 involve LLMs as any important, original, or non-standard components.
- 957 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
958 for what should or should not be described.