A Proofs

The following result strengthens Proposition 1 and provides a sufficient condition under which f and its convex envelope f_c have the same set of minimizers. This result implies that one can minimize the function f by minimizing its convex envelope f_c, under the assumption that the set of minimizer of f, $\mathcal{X}_f^*$, is a convex set.

Lemma 2. *Let f_c be the convex envelope of f on $\mathcal{X}$. Let $\mathcal{X}_{f_c}^*$ be the set of minimizers of f_c. Assume that $\mathcal{X}_f^*$ is a convex set. Then $\mathcal{X}_{f_c}^* = \mathcal{X}_f^*$.*

Proof. We prove this result by a contradiction argument. Assume that the result is not true. Then there exists some $\widetilde{x} \in \mathcal{X}$ such that $f_c(\widetilde{x}) = f^*$ and $\widetilde{x} \notin \mathcal{X}_f^*$, i.e., $f(\widetilde{x}) > f^*$. By definition of the convex envelope, $(f^*, \widetilde{x})$ lies in $\mathrm{conv}(\mathrm{epi}f)$. This combined with the fact that $\mathrm{conv}(\mathrm{epi}f)$ is the smallest convex set which contains $\mathrm{epi}f$, implies that there exists some $z_1 = (\xi_1, x_1)$ and $z_2 = (\xi_2, x_2)$ in $\mathrm{epi}f$ and $0 \leq \alpha \leq 1$ such that

$$(f^*, \widetilde{x}) = \alpha z_1 + (1-\alpha) z_2. \tag{6}$$

Let us first consider the case in which z_1 and z_2 belong to the set $\widetilde{\mathcal{X}^*} = \{(\xi, x) | x \in \mathcal{X}_f^*, \xi = f(x)\}$. The set $\widetilde{\mathcal{X}^*}$ is convex. So every convex combination of its entries also belongs to $\widetilde{\mathcal{X}^*}$ as well. This is not the case for z_1 and z_2 due to the fact that $(f^*, \widetilde{x}) = \alpha z_1 + (1-\alpha) z_2$ does not belong to $\widetilde{\mathcal{X}^*}$ as $\widetilde{x} \notin \mathcal{X}^*$. Now consider the case that either z_1 or z_2 are not in $\widetilde{\mathcal{X}^*}$. Without loss of generality, assume that $z_1 \notin \widetilde{\mathcal{X}^*}$. In this case, ξ_1 must be larger than f^* since $x_1 \notin \mathcal{X}_f^*$. This implies that $(f^*, \widetilde{x})$ can not be expressed as the convex combination of z_1 and z_2 since in this case: **(i)** for every $0 < \alpha \leq 1$, we have that $\alpha\xi_1 + (1-\alpha)\xi_2 > f^*$ and **(ii)** when $\alpha = 0$, then $x_2 = \widetilde{x}$ and therefore $\alpha\xi_1 + (1-\alpha)\xi_2 = \xi_2 = f(\widetilde{x}) > f^*$. Therefore Eqn. 6 can not hold for any $z_1, z_2 \in \mathrm{epi}f$ when $0 \leq \alpha \leq 1$. Thus the assumption that there exists some $\widetilde{x} \in \mathcal{X}/\mathcal{X}_f^*$ such that $f_c(\widetilde{x}) = f^*$ can not be true either, which proves the result. □

A.1 Proof of Lem. 1

We first prove that any underestimate (lower bound) of function f (except f_c) does not satisfy the constraint of the optimization problem of Eqn. 2. This is due to the fact that for any underestimate $h(\cdot;\theta) \in \mathcal{H}/f_c$, there exists some $x_u \in \mathcal{X}$ and $\varepsilon > 0$ such that for every $\theta_c \in \Theta_c$

$$\begin{aligned} |h(x_u;\theta) - h(x_u;\theta_c)| &= h(x_u;\theta_c) - h(x_u;\theta) \\ &= f_c(x_u) - h(x_u;\theta) = \varepsilon. \end{aligned}$$

For every $x \in \mathcal{X}$, the following then holds due to the fact that the function class $\mathcal{H}$ is assumed to be Lipschitz:

$$\begin{aligned} &h(x;\theta) - h(x;\theta_c) = h(x;\theta) - h(x_u,\theta) - \varepsilon \\ &h(x_u,\theta_c) - h(x;\theta_c) \leq 2\lambda d(x, x_u) - \varepsilon. \end{aligned} \tag{7}$$

Eqn. 7 implies that for every $x \in \mathcal{B}(x_u, \varepsilon/2\lambda)$ the inequality $\Delta_c(x) = h(x;\theta_c) - h(x;\theta) > 0$ holds. Denote the event $\{x \in \mathcal{B}(x_u, \varepsilon/(2\lambda))\}$ by Ω_u. We then deduce that

$$\mathbb{E}[\Delta_c(x)] \geq \mathbb{P}(\Omega_u)\mathbb{E}[\Delta_c(x)|\Omega_u] > 0,$$

where the last inequality follows due to the fact that both $\mathbb{P}(\Omega_u)$ and $\mathbb{E}[\Delta_c(x)|\Omega_u]$ are larger than 0. The inequality $\mathbb{P}(\Omega_u) > 0$ holds since $\rho(x) > 0$ for every $x \in \mathcal{X}$ and also that $\mathcal{B}(x_u, \varepsilon/2\lambda) \neq \emptyset$. The inequality $\mathbb{E}[\Delta_c(x)|\Omega_u] > 0$ holds by the fact that for every $x \in \mathcal{B}(x_u, \varepsilon/2\lambda)$ the inequality $\Delta_c(x) > 0$ holds.

Let $\widetilde{\mathcal{H}} := \{h : h \in \mathcal{H}, \mathbb{E}[h(x;\theta)] = \mathbb{E}[f_c(x)]\}$ be a set of all functions h in $\mathcal{H}$ with the same mean as the convex envelope f_c. We now show that f_c is the only minimizer of $L(\theta) = \mathbb{E}[|h(x;\theta) - f(x)|]$ that lies in the set $\widetilde{\mathcal{H}}$. We do this by proving that for every $h \in \widetilde{\mathcal{H}}/f_c$, the loss $L(\theta) > L(\theta_c)$, for every $\theta_c \in \Theta_c$. First we recall that any underestimate $h \in \mathcal{H}/f_c$ of f can not lie in $\widetilde{\mathcal{H}}$, as we have already shown that $\mathbb{E}[h(x;\theta)] < \mathbb{E}[f_c(x)]$ for every $h \in \mathcal{H}/f_c$. This implies that for every $h \in \widetilde{\mathcal{H}}/f_c$ there exists some $x_o \in \mathcal{X}$ such that $h(x_o;\theta) > f(x)$, or equivalently, we have that for every $h \in \widetilde{\mathcal{H}}/f_c$ there exists some $x_o \in \mathcal{X}$ and $\varepsilon > 0$ such that

$$|h(x_o;\theta) - f(x_o)| = h(x_o;\theta) - f(x_o) = \varepsilon.$$

Then for every $x \in \mathcal{X}$, the following holds due to the fact that the function class $\mathcal{H}$ and f are assumed to be Lipschitz:

$$h(x;\theta) - f(x) = h(x;\theta) - h(x_o,\theta) + \varepsilon \tag{8}$$
$$f(x_o) - f(x) \geq -2\lambda d(x, x_o). \tag{9}$$

Eqn. 8 implies that for every $x \in \mathcal{B}(x_o, \varepsilon/2\lambda)$ the inequality $h(x;\theta) - f_c(x) > 0$ holds. Denote the event $\{x \in \mathcal{B}(x_o, \varepsilon/2\lambda)\}$ by Ω_o. Let $\Delta(x) = f(x) - h(x;\theta)$. We then deduce

$$\begin{aligned} &\mathbb{E}[|h(x;\theta) - f(x)|] \\ &= \mathbb{P}(\Omega_o)\mathbb{E}[|\Delta(x)| \mid \Omega_o] + \mathbb{P}(\Omega_o^c)\mathbb{E}[|\Delta(x)| \mid \Omega_o^c] \qquad (10) \\ &> \mathbb{P}(\Omega_o)\mathbb{E}[\Delta(x) \mid \Omega_o] + \mathbb{P}(\Omega_o^c)\mathbb{E}[\Delta(x) \mid \Omega_o^c] \qquad (11) \\ &= \mathbb{E}[\Delta(x)] = \mathbb{E}[f(x) - f_c(x)]. \qquad (12) \end{aligned}$$

Line (10) holds by the law of total expectation. The inequality (11) holds since $h(x;\theta) > f(x)$ for every $x \in \mathcal{B}(x_o, \varepsilon/2\lambda)$. This implies that $|h(x;\theta) - f(x)| > 0 > f(x) - h(x;\theta)$. Line (12) holds since $\mathbb{E}[h(x;\theta)] = \mathbb{E}[f_c(x)]$ for $h \in \widetilde{\mathcal{H}}$. The fact that $L(\theta) = \mathbb{E}[|h(x;\theta) - f(x)|] > \mathbb{E}[|f(x) - f_c(x)|] = L(\theta_c)$ for every $h(\cdot;\theta) \in \mathcal{H}/f_c$ implies that the set of minimizers of $L(\theta)$ coincide with the set Θ_c, which completes the proof.

A.2 Proof of Thm. 1

To prove the result of Thm. 1, we need to relate the solution of the optimization problem of Eqn. 4 with the result of Alg. 1, for which we rely on the following lemmas.

Before we proceed, we must introduce some new notation. Define the convex sets Θ^e and $\widehat{\Theta}^e$ as $\Theta^e := \{\theta : \theta \in \Theta, \mathbb{E}[h(x;\theta)] = \mathbb{E}[f_c(x)]\}$ and $\widehat{\Theta}^e := \{\theta : \theta \in \Theta, \widehat{\mathbb{E}}_2[h(x;\theta)] = \widehat{\mathbb{E}}_2[f_c(x)]\}$, respectively. Also define the subspace $\Theta_{\text{sub}} := \{\theta : \theta \in \mathbb{R}^p, \mathbb{E}[h(x;\theta)] = \mathbb{E}[f_c(x)]\}$.

Lemma 3. *Let δ be a positive scalar. Under Assumptions 1 and 3 there exists some $\mu \in [-R, R]$ such that the following holds w.p. $1-\delta$:*

$$\left|L(\widehat{\theta}_\mu) - \min_{\theta\in\Theta^e} L(\theta)\right| \leq \mathcal{O}\left(BRU\sqrt{\frac{\log(1/\delta)}{T}}\right).$$

Proof. The empirical estimate $\widehat{\theta}_\mu$ is obtained by minimizing the empirical $\widehat{L}(\theta)$ under some affine constraints. Additionally, the function $L(\theta)$ takes the form of the expected value of a generalized linear model. Now set $\mu = \widehat{\mathbb{E}}_2[f_c(x)]$. In this case, the following result on stochastic optimization of the generalized linear model holds for $\mu = \widehat{\mathbb{E}}_2[f_c(x)]$ w.p. $1-\delta$ (see, e.g., Shalev-Shwartz et al., 2009, for the proof):

$$L(\widehat{\theta}_\mu) - \min_{\theta\in\widehat{\Theta}^e} L(\theta) = \mathcal{O}\left(BRU_1\sqrt{\frac{\log(1/\delta)}{T}}\right),$$

where U_1 is the Lipschitz constant of $|h(x;\theta) - f(x)|$. We then deduce that for every $x \in \mathcal{X}$, $\theta \in \Theta$ and $\theta' \in \Theta$,

$$|\,|h(x,\theta) - f(x)| - |h(x,\theta') - f(x)|\,| \leq U_1\|\theta - \theta'\|.$$

The inequality $|\,|a| - |b|\,| \leq |a-b|$, combined with the fact that for every $x \in \mathcal{X}$ the function $h(x;\theta)$ is Lipschitz continuous in θ implies,

$$\begin{aligned}&|\,|h(x,\theta) - f(x)| - |h(x,\theta') - f(x)|\,| \\ \leq&|h(x,\theta) - h(x,\theta')| \leq U\|\theta - \theta'\|.\end{aligned}$$

Therefore the following holds:

$$L(\widehat{\theta}_\mu) - \min_{\theta\in\widehat{\Theta}^e} L(\theta) = \mathcal{O}\left(BRU\sqrt{\frac{\log(1/\delta)}{T}}\right). \quad (13)$$

For every $\theta \in \widehat{\Theta}^e$, the following holds w.p. $1-\delta$:

$$\begin{aligned}\mathbb{E}[h(x;\theta)] - \widehat{\mathbb{E}}_2[f_c(x)] &= \mathbb{E}[h(x;\theta)] - \widehat{\mathbb{E}}_2[h(x;\theta)] \\ &\leq R\sqrt{\frac{\log(1/\delta)}{2T}},\end{aligned}$$

as well as,

$$\widehat{\mathbb{E}}_2[f_c(x)] - \mathbb{E}[f_c(x)] \leq R\sqrt{\frac{\log(1/\delta)}{2T}},$$

in which we rely on the Höeffding inequality for concentration of measure. These results combined with a union bound argument implies that:

$$\begin{aligned}\mathbb{E}[h(x;\theta)] - \mathbb{E}[f_c(x)] &= \mathbb{E}[h(x;\theta)] - \widehat{\mathbb{E}}_2[f_c(x)] \\ &\quad + \widehat{\mathbb{E}}_2[f_c(x)] - \mathbb{E}[f_c(x)] \\ &\leq R\sqrt{\frac{2\log(2/\delta)}{T}},\end{aligned} \quad (14)$$

for every $\theta \in \widehat{\Theta}^e$. We know that $\min_{\theta\in\widehat{\Theta}^e} L(\theta) \leq L(\theta_c)$, due the fact that $\theta_c \in \widehat{\Theta}^e$. This combined with the fact that $\theta_c = \min_{\theta\in\Theta^e} L(\theta)$ leads to the following sequence of inequalities w.p. $1-\delta$:

$$\begin{aligned}&\min_{\theta\in\widehat{\Theta}^e} L(\theta) \leq L(\theta_c) = \mathbb{E}[f(x) - f_c(x)] \\ &\leq \mathbb{E}[|f(x) - h(x;\widehat{\theta}_c)|] + \mathbb{E}[h(x;\widehat{\theta}_c) - f_c(x)] \\ &\leq \min_{\theta\in\widehat{\Theta}^e} L(\theta) + R\sqrt{\frac{2\log(2/\delta)}{T}},\end{aligned}$$

where the last inequality follows from the bound of Eqn. 14. It immediately follows that:

$$\left|\min_{\theta\in\widehat{\Theta}^e} L(\theta) - \min_{\theta\in\Theta^e} L(\theta)\right| \leq R\sqrt{\frac{2\log(2/\delta)}{T}},$$

w.p. $1-\delta$. This combined with Eqn. 13 completes the proof. $\square$

Let $\widehat{\theta}_\mu^{\text{proj}}$ be the ℓ_2-normed projection of $\widehat{\theta}_\mu$ on the subspace Θ_{sub}. We now prove bound on the error $\|\widehat{\theta}_\mu^{\text{proj}} - \widehat{\theta}_\mu\|$.

Lemma 4. *Let δ be a positive scalar. Then under Assumptions 1 and 3 there exists some $\mu \in [-R, R]$ such that the following holds with probability $1-\delta$:*

$$\|\widehat{\theta}_\mu^{\text{proj}} - \widehat{\theta}_\mu\| \leq \frac{R}{\|\mathbb{E}[\phi(x)]\|}\sqrt{\frac{2\log(4/\delta)}{T}}.$$

Proof. Set $\mu = \mu_f := \mathbb{E}[f_c(x)]$. Then $\widehat{\theta}_\mu^{\text{proj}}$ can be obtained as the solution of following optimization problem:

$$\widehat{\theta}_\mu^{\text{proj}} = \arg\min_{\theta\in\mathbb{R}^p}\|\theta - \widehat{\theta}_\mu\|^2 \qquad \text{s.t.} \qquad \mathbb{E}[h(x;\theta)] = \mu_f.$$

Thus $\widehat{\theta}_\mu^{\text{proj}}$ can be obtain as the extremum of the following Lagrangian:

$$\mathcal{L}(\theta,\lambda) = \|\theta - \widehat{\theta}_\mu\|^2 + \lambda(\mathbb{E}[h(x;\theta)] - \mu_f).$$

This problem can be solved in closed-form as follows:

$$\begin{aligned}0 &= \frac{\partial\mathcal{L}(\theta,\lambda)}{\partial\theta} = \theta - \widehat{\theta}_\mu + \lambda\mathbb{E}[\phi(x)] \\ 0 &= \frac{\partial\mathcal{L}(\theta,\lambda)}{\partial\lambda} = \mathbb{E}[h(x;\theta)] - \mu_f.\end{aligned} \quad (15)$$

Solving the above system of equations leads to $\mathbb{E}[h(x;(\widehat{\theta}_\mu - \lambda\mathbb{E}[\phi(x)])] = \mu_f$. The solution for λ can be obtained as

$$\lambda = \frac{\mu_f - \mathbb{E}[h(x;\widehat{\theta}_\mu)]}{\|\mathbb{E}[\phi(x)]\|^2}.$$

By plugging this in Eqn. 15 we deduce:

$$\widehat{\theta}^{\mathrm{proj}}_\mu = \widehat{\theta}_\mu - \frac{(\mu_f - \mathbb{E}[h(x;\widehat{\theta}_\mu)])\mathbb{E}[\phi(x)]}{\|\mathbb{E}[\phi(x)]\|^2},$$

For the choice of $\mu = \widehat{\mathbb{E}}_2[f_c(x)]$ we deduce:

$$\begin{aligned}\|\widehat{\theta}^{\mathrm{proj}}_\mu - \widehat{\theta}_\mu\| &= \frac{|\mu_f - \mathbb{E}[h(x;\widehat{\theta}_\mu)]|}{\|\mathbb{E}[\phi(x)]\|} \\ &= \frac{|\mathbb{E}[f_c(x)] - \mathbb{E}[h(x;\widehat{\theta}_\mu)]|}{\|\mathbb{E}[\phi(x)]\|}.\end{aligned}$$

This combined with Eqn. 14 and a union bound proves the result. □

We proceed by proving bound on the absolute error $|L(\widehat{\theta}^{\mathrm{proj}}_\mu) - L(\theta_c)| = |L(\widehat{\theta}^{\mathrm{proj}}_\mu) - \min_{\theta\in\Theta^e} L(\theta)|$.

Lemma 5. *Let δ be a positive scalar. Under Assumptions 1 and 3 there exists some $\mu \in [-R, R]$ such that the following holds with probability $1-\delta$:*

$$\left|L(\widehat{\theta}^{\mathrm{proj}}_\mu) - L(\theta_c)\right| = \mathcal{O}\left(BRU\sqrt{\frac{\log(1/\delta)}{T}}\right).$$

Proof. From Lem. 4 we deduce:

$$\begin{aligned}&|\mathbb{E}[h(x;\widehat{\theta}^{\mathrm{proj}}_\mu) - h(x;\widehat{\theta}_\mu)]| \\ &\le \|\widehat{\theta}^{\mathrm{proj}}_\mu - \widehat{\theta}_\mu\|\|\mathbb{E}[\phi(x)]\| \le 2R\sqrt{\frac{\log(4/\delta)}{T}},\end{aligned} \tag{16}$$

where the first inequality is due to the Cauchy-Schwarz inequality. We then deduce:

$$\begin{aligned}&|\,|L(\widehat{\theta}^{\mathrm{proj}}_\mu) - L(\theta_c)| - |L(\widehat{\theta}_\mu) - L(\theta_c)|\,| \\ &\le |L(\widehat{\theta}^{\mathrm{proj}}_\mu) - L(\widehat{\theta}_\mu)| \le |\mathbb{E}[h(x;\widehat{\theta}^{\mathrm{proj}}_\mu) - h(x;\widehat{\theta}_\mu)]|,\end{aligned}$$

in which we rely on the triangle inequality $|\,|a| - |b|\,| \le |a-b|$. It then follows that

$$\begin{aligned}L(\widetilde{\theta}_\mu) - L(\theta_c) \le{}& |L(\widehat{\theta}_\mu) - L(\theta_c)| \\ &+ |\mathbb{E}[h(x;\widehat{\theta}^{\mathrm{proj}}_\mu) - h(x;\widehat{\theta}_\mu)]|.\end{aligned}$$

Combining this result with the result of Lem. 3 and Eqn. 16 proves the result.

□

In the following lemma we make use of Lem. 4 and Lem. 5 to prove that the minimizer $\widehat{x}_\mu = \arg\min_{x\in\mathcal{X}} h(x;\widehat{\theta}_\mu)$ is close to a global minimizer $x^* \in \mathcal{X}^*_f$.

Lemma 6. *Under Assumptions 1, 3 and 4 there exists some $\mu \in [-R, R]$ such that w.p. $1-\delta$:*

$$d(\widehat{x}_\mu, \mathcal{X}^*_f) = \mathcal{O}\left(\left(\frac{\log(1/\delta)}{T}\right)^{\beta_1\beta_2/2}\right).$$

Proof. The result of Lem. 5 combined with Assumption 4.b implies that w.p. $1-\delta$:

$$d_2(\widehat{\theta}^{\mathrm{proj}}_\mu, \Theta_c) \le \left(\frac{\varepsilon_1(\delta)}{\gamma}\right)^{\beta_2},$$

where $\varepsilon_1(\delta) = BRU\sqrt{\frac{\log(1/\delta)}{T}}$. This combined with the result of Lem. 4 implies that w.p. $1-\delta$:

$$d_2(\widehat{\theta}_\mu, \Theta_c) \le d_2(\widehat{\theta}^{\mathrm{proj}}_\mu, \Theta_c) + d_2(\widehat{\theta}^{\mathrm{proj}}_\mu, \widehat{\theta}_\mu) \le 2\left(\frac{\varepsilon_c(\delta)}{\gamma_2}\right)^{\beta_2},$$

where $\varepsilon_c(\delta) = \mathcal{O}\left(\frac{RBU}{\min(1,\|\mathbb{E}[\phi(x)]\|)}\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right)$.

We now use this result to prove a high probability bound on $f_c(\widehat{x}_\mu) - f^*$:

$$\begin{aligned}&f_c(\widehat{x}_\mu) - f^* = h(\theta_c, \widehat{x}_\mu) - h(\theta_c, x^*) \\ &= h(\theta_c, \widehat{x}_\mu) - h(\widehat{\theta}_\mu, \widehat{x}_\mu) + \min_{x\in\mathcal{X}} h(\widehat{\theta}_\mu, x) - h(\theta_c, x^*) \\ &\le h(\theta_c, \widehat{x}_\mu) - h(\widehat{\theta}_\mu, \widehat{x}_\mu) + h(\widehat{\theta}_\mu, x^*) - h(\theta_c, x^*) \\ &\le 2U d_2(\widehat{\theta}_\mu, \Theta_c) \le 2U\left(\frac{\varepsilon_c(\delta)}{\gamma_2}\right)^{\beta_2},\end{aligned}$$

where the last inequality follows by the fact that h is U-Lipschitz w.r.t. θ. This combined with Assumption 4.a completes the proof.

□

It then follows by combining the result of Lem. 6, Assumption 2 and the fact that f_c is the tightest convex lower bound of function f that there exist a $\mu = [-R, R]$ such that

$$f(\widehat{x}_\mu) - f^* = \mathcal{O}\left[\left(\frac{\log(1/\delta)}{T}\right)^{\beta_1\beta_2/2}\right]$$

This combined with the fact that $f(\widehat{x}_{\widehat{\mu}}) \le f(\widehat{x}_\mu)$ for every $\mu \in [-R, R]$, completes the proof of the main result (Thm. 1) .

A.3 Proof of Thm. 2

We prove this theorem by generalizing the result of Lems. 3-6 to the case that $f \notin \mathcal{H}$. First we need to introduce some notation. Under the assumptions of Thm. 2, for every $\zeta > 0$, there exists some $\theta^\zeta \in \Theta$ and $\upsilon > 0$ such that the following inequality holds:

$$\mathbb{E}[|h(x;\theta^\zeta) - f_c(x)|] \leq \upsilon + \zeta.$$

Define the convex sets $\widetilde{\Theta}^\zeta := \{\theta : \theta \in \Theta, \mathbb{E}_2[h(x;\theta)] = \mathbb{E}_2[h(x;\theta^\zeta)]\}$ and $\widehat{\Theta}^\zeta := \{\theta : \theta \in \Theta, \widehat{\mathbb{E}}_2[h(x;\theta)] = \widehat{\mathbb{E}}_2[h(x;\theta^\zeta)]\}$. Also define the subspace $\Theta^\zeta_{\text{sub}} := \{\theta : \theta \in \mathbb{R}^{\widetilde{p}}, \mathbb{E}[h(x;\theta)] = \mathbb{E}[h(x;\theta^\zeta)]\}$.

Lemma 7. *Let δ be a positive scalar. Under Assumptions 1 and 5 there exists some $\mu \in [-R, R]$ such that for every $\zeta > 0$ the following holds with probability $1-\delta$:*

$$\left|L(\widehat{\theta}_\mu) - \min_{\theta \in \widetilde{\Theta}^\zeta} L(\theta)\right| = \mathcal{O}\left(BRU\sqrt{\frac{\log(1/\delta)}{T}}\right) + \upsilon + \zeta.$$

Proof. The empirical estimate $\widehat{\theta}_\mu$ is obtained by minimizing the empirical $\widehat{L}(\theta)$ under some affine constraints. Also the function $L(\theta)$ is in the form of expected value of some generalized linear model. Now set $\mu = \widehat{\mathbb{E}}_2[h(x;\theta^\zeta)]$. Then the following result on stochastic optimization of the generalized linear model holds w.p. $1-\delta$ (see, e.g., Shalev-Shwartz et al., 2009, for the proof):

$$L(\widehat{\theta}_\mu) - \min_{\theta \in \widehat{\Theta}^\zeta} L(\theta) = \mathcal{O}\left(BRU_1\sqrt{\frac{\log(1/\delta)}{T}}\right),$$

where U_1 satisfies the following Lipschitz continuity inequality for every $x \in \mathcal{X}$, $\theta \in \Theta$ and $\theta' \in \Theta$:

$$|\,|h(x,\theta) - f(x)| - |h(x,\theta') - f(x)|\,| \leq U_1\|\theta - \theta'\|.$$

The inequality $|\,|a| - |b|\,| \leq |a-b|$ combined with the fact that for every $x \in \mathcal{X}$ the function $h(x;\theta)$ is Lipschitz continuous in θ implies

$$\begin{aligned}&|\,|h(x,\theta) - f(x)| - |h(x,\theta') - f(x)|\,| \\ \leq &|h(x,\theta) - h(x,\theta')| \leq U\|\theta - \theta'\|.\end{aligned}$$

Therefore the following holds:

$$L(\widehat{\theta}_\mu) - \min_{\theta \in \widehat{\Theta}^\zeta} L(\theta) = \mathcal{O}\left(BRU\sqrt{\frac{\log(1/\delta)}{T}}\right), \quad (17)$$

For every $\theta \in \widehat{\Theta}^\zeta$ the following holds w.p. $1-\delta$:

$$\begin{aligned}&\mathbb{E}[h(x;\theta)] - \widehat{\mathbb{E}}_2[h(x;\theta^\zeta)] = \mathbb{E}[h(x;\theta)] - \widehat{\mathbb{E}}_2[h(x;\theta)] \\ &\leq R\sqrt{\frac{\log(1/\delta)}{2T}},\end{aligned}$$

as well as,

$$\widehat{\mathbb{E}}_2[h(x;\theta^\zeta)] - \mathbb{E}[h(x;\theta^\zeta)] \leq R\sqrt{\frac{\log(1/\delta)}{2T}},$$

in which we rely on the Höeffding inequality for concentration of measure. These results combined with a union bound argument implies that

$$\begin{aligned}&\mathbb{E}[h(x;\theta)] - \mathbb{E}[h(x;\theta^\zeta)] = \mathbb{E}[h(x;\theta)] - \widehat{\mathbb{E}}_2[h(x;\theta^\zeta)] \\ &+ \widehat{\mathbb{E}}_2[h(x;\theta^\zeta)] - \mathbb{E}[h(x;\theta^\zeta)] \leq R\sqrt{\frac{2\log(2/\delta)}{T}},\end{aligned} \quad (18)$$

for every $\theta \in \widehat{\Theta}^\zeta$. Then the following sequence of inequalities holds:

$$\begin{aligned}&\min_{\theta \in \widehat{\Theta}^\zeta} L(\theta) \leq L(\theta^\zeta) = \mathbb{E}[|h(x;\theta^\zeta) - f(x)|] \\ \leq &L(\theta_c) + \mathbb{E}[|h(x;\theta^\zeta) - f_c(x)|] \\ \leq &L(\theta_c) + \upsilon + \zeta \\ \leq &\min_{\theta \in \widehat{\Theta}^\zeta} L(\theta) + R\sqrt{\frac{2\log(2/\delta)}{T}}.\end{aligned}$$

The first inequality follows from the fact that $\theta_c \in \widehat{\Theta}^\zeta$. Also the following holds w.p. $1-\delta$:

$$\begin{aligned}&L(\theta_c) \leq \mathbb{E}[|h(x;\theta^\zeta) - f_c(x)|] + \mathbb{E}[h(x;\theta^\zeta)] - \mathbb{E}[f(x)] \\ \leq &\upsilon + \zeta + \mathbb{E}[h(x;\theta^\zeta)] - \mathbb{E}[f(x)] \\ \leq &\min_{\theta \in \widehat{\Theta}^\zeta} \mathbb{E}[h(x;\theta)] - \mathbb{E}[f(x)] + R\sqrt{\frac{2\log(2/\delta)}{T}} + \upsilon + \zeta \\ \leq &\min_{\theta \in \widehat{\Theta}^\zeta} L(\theta) + R\sqrt{\frac{2\log(2/\delta)}{T}} + \upsilon + \zeta.\end{aligned}$$

The last inequality follows from the bound of Eqn. 18. It immediately follows that

$$\left|\min_{\theta \in \widehat{\Theta}^\zeta} L(\theta) - \min_{\theta \in \Theta^e} L(\theta)\right| \leq R\sqrt{\frac{2\log(2/\delta)}{T}} + \upsilon + \zeta,$$

w.p. $1-\delta$. This combined with Eqn. 17 completes the proof.

$\square$

Under Assumption 6, for every $h(\cdot;\theta) \in \mathcal{H}$, there exists some $h(\cdot;\widetilde{\theta}) \in \widetilde{\mathcal{H}}$ such that $h(x;\theta) = h(x;\widetilde{\theta})$ for every $x \in \mathcal{X}$. Let $\widetilde{\theta}_\mu$ be the corresponding set of parameters for $\widehat{\theta}_\mu$ in $\widetilde{\Theta}$. Let $\widetilde{\theta}^{\text{proj}}_\mu$ be the ℓ_2-normed projection of $\widetilde{\theta}_\mu$ on the subspace $\Theta^\zeta_{\text{sub}}$. We now prove bound on the error $\|\widetilde{\theta}_\mu - \widetilde{\theta}^{\text{proj}}_\mu\|$.

Lemma 8. *Under Assumptions 1 and 5 and 6 there exists some $\mu \in [-R, R]$ such that the following holds with probability $1-\delta$:*

$$\|\widetilde{\theta}_\mu^{\text{proj}} - \widetilde{\theta}_\mu\| \leq \frac{R\sqrt{\frac{2\log(4/\delta)}{T}} + \upsilon + \zeta}{\|\mathbb{E}[\phi(x)]\|},$$

Proof. $\widetilde{\theta}_\mu^{\text{proj}}$ is the solution of following optimization problem:

$$\widetilde{\theta}_\mu^{\text{proj}} = \arg\min_{\theta \in \mathbb{R}^{\widetilde{p}}} \|\theta - \widehat{\theta}_\mu\|^2 \qquad \text{s.t.} \qquad \mathbb{E}[h(x;\theta)] = \mu_f,$$

where $\mu_f = \mathbb{E}[f_c(x)]$. Thus $\widetilde{\theta}_\mu^{\text{proj}}$ can be obtain as the extremum of the following Lagrangian:

$$\mathcal{L}(\theta, \lambda) = \|\theta - \widetilde{\theta}_\mu\|^2 + \lambda(\mathbb{E}[h(x;\theta)] - \mu_f).$$

This problem can be solved in closed-form as follows:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta} = \theta - \widetilde{\theta}_\mu + \lambda \mathbb{E}[(\widetilde{\phi}(x)] \qquad (19)\\ 0 &= \frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \lambda} = \mathbb{E}[h(x;\theta)] - \mu_f. \end{aligned}$$

Solving the above system of equations leads to $\mathbb{E}[h(x;\widetilde{\theta}_\mu)] - \lambda \mathbb{E}[\widetilde{\phi}(x)] = \mu_f$. The solution for λ can be obtained as

$$\lambda = \frac{\mu - \mathbb{E}[h(x;\widetilde{\theta}_\mu)]}{\|\mathbb{E}[\widetilde{\phi}(x)]\|^2}.$$

By plugging this in Eqn. 19 we deduce:

$$\widetilde{\theta}_\mu^{\text{proj}} = \widetilde{\theta}_\mu - \frac{(\mu_f - \mathbb{E}[h(x;\widetilde{\theta}_\mu)])\mathbb{E}[\widetilde{\phi}(x)]}{\|\mathbb{E}[\widetilde{\phi}(x)]\|^2},$$

We then deduce:

$$\begin{aligned} &\|\widetilde{\theta}_\mu^{\text{proj}} - \widetilde{\theta}_\mu\| = \frac{|\mu_f - \mathbb{E}[h(x;\widehat{\theta}_\mu)]|}{\|\mathbb{E}[\widetilde{\phi}(x)]\|} \\ \leq &\frac{\mathbb{E}[|f_c(x) - h(x;\theta^\zeta)|] + |\mathbb{E}[h(x;\theta^\zeta)] - \mathbb{E}[h(x;\widehat{\theta}_\mu)]|}{\|\mathbb{E}[\widetilde{\phi}(x)]\|}. \end{aligned}$$

This combined with Eqn. 18 and a union bound proves the result. ☐

We proceed by proving bound on the absolute error $|L(\widetilde{\theta}_\mu^{\text{proj}}) - L(\theta_c)| = |L(\widetilde{\theta}_\mu^{\text{proj}}) - \min_{\theta \in \widetilde{\Theta}} L(\theta)|$.

Lemma 9. *Under Assumptions 1, 5 and 6 there exists some $\mu \in [-R, R]$ such that for every $\zeta > 0$ the following bound holds with probability $1-\delta$:*

$$\left|L(\widetilde{\theta}_\mu^{\text{proj}}) - L(\theta_c)\right| = \mathcal{O}\left(\zeta + \upsilon + BRU\sqrt{\frac{\log(1/\delta)}{T}}\right).$$

Proof. From Lem. 8 we deduce

$$\begin{aligned} &|\mathbb{E}[h(x;\widetilde{\theta}_\mu^{\text{proj}}) - h(x;\widetilde{\theta}_\mu)]| \\ &\leq \|\widetilde{\theta}_\mu^{\text{proj}} - \widetilde{\theta}_\mu\|\|\mathbb{E}[\widetilde{\phi}(x)]\| \leq 2R\sqrt{\frac{\log(4/\delta)}{T}} + \zeta + \upsilon. \end{aligned} \tag{20}$$

where in the first inequality we rely on the Cauchy-Schwarz inequality. We then deduce:

$$\begin{aligned} &|\,|L(\widetilde{\theta}_\mu^{\text{proj}}) - L(\theta_c)| - |L(\widetilde{\theta}_\mu) - L(\theta_c)|\,| \\ &\leq |L(\widetilde{\theta}_\mu^{\text{proj}}) - L(\widetilde{\theta}_\mu)| \leq |\mathbb{E}[h(x;\widetilde{\theta}_\mu^{\text{proj}}) - h(x;\widehat{\theta}_\mu)]|, \end{aligned}$$

in which we rely on the triangle inequality $|\,|a| - |b|\,| \leq |a-b|$. We then deduce

$$\begin{aligned} L(\widetilde{\theta}_\mu^{\text{proj}}) - L(\theta_c) \leq &|L(\widehat{\theta}_\mu) - L(\theta_c)| \\ &+ |\mathbb{E}[h(x;\widetilde{\theta}_\mu^{\text{proj}}) - h(x;\widetilde{\theta}_\mu)]|. \end{aligned}$$

Combining this result with the result of Lem. 7 and Eqn. 20 proves the main result.

☐

In the following lemma, we make use of Lem. 8 and Lem. 9 to prove that the minimizer $\widehat{x}_\mu = \arg\min_{x \in \mathcal{X}} h(x;\widehat{\theta}_\mu)$ is near a global minimizer $x^* \in \mathcal{X}_f^*$ w.r.t. to the metric d.

Lemma 10. *Under Assumptions 1, 5 and 6 there exists some $\mu \in [-R, R]$ such that w.p. $1-\delta$:*

$$d(\widehat{x}_\mu, \mathcal{X}_f^*) = \mathcal{O}\left[\left(\sqrt{\frac{\log(1/\delta)}{T}} + \zeta + \upsilon\right)^{\beta_1\beta_2}\right].$$

Proof. The result of Lem. 9 combined with Assumption 6.b implies that w.p. $1-\delta$:

$$d_2(\theta_\mu^{\text{proj}}, \Theta_c) \leq \left(\frac{\varepsilon_1(\theta)}{\gamma_2}\right)^{\beta_2},$$

where $\varepsilon_1(\theta) = \mathcal{O}(BRU\sqrt{\frac{\log(1/\delta)}{T}} + \upsilon + \zeta)$. This combined with the result of Lem. 8 implies that w.p. $1-\delta$:

$$d_2(\widetilde{\theta}_\mu, \theta_c) \leq d_2(\widetilde{\theta}_\mu^{\text{proj}}, \theta_c) + d_2(\widetilde{\theta}_\mu^{\text{proj}}, \widetilde{\theta}_\mu) \leq 2\left(\frac{\varepsilon_c(\delta)}{\gamma_2}\right)^{\beta_2},$$

where $\varepsilon_c(\delta)$ is defined as:

$$\varepsilon_c(\delta) := \mathcal{O}\left(\frac{RBU\sqrt{\frac{\log(1/\delta)}{T}} + \zeta + \upsilon}{\min(1, \|\mathbb{E}[\widetilde{\phi}(x)))\|]}\right).$$

We now use this result to prove high probability bound on $f_c(\widehat{x}_\mu) - f^*$:

$$\begin{aligned}
&f_c(\widehat{x}_\mu) - f^* = h(\theta_c, \widehat{x}_\mu) - h(\theta_c, x^*) \\
&= h(\theta_c, \widehat{x}_\mu) - h(\widehat{\theta}_\mu, \widehat{x}_\mu) + \min_{x \in \mathcal{X}} h(\widehat{\theta}_\mu, x) - h(\theta_c, x^*) \\
&\leq h(\theta_c, \widehat{x}_\mu) - h(\widehat{\theta}_\mu, \widehat{x}_\mu) + h(\widehat{\theta}_\mu, x^*) - h(\theta_c, x^*) \\
&\leq 2U d_2(\widehat{\theta}_\mu, \Theta_c) \leq 2\gamma_2 U \left(\frac{\varepsilon_c(\delta)}{\gamma_2}\right)^{\beta_2},
\end{aligned}$$

where the last inequality follows by the fact that h is U-Lipschitz w.r.t. θ. This combined with Assumption 6.a completes the proof.

□

It then follows by combining the result of Lem. 10 and Assumption 2 that there exist a $\mu \in [-R, R]$ such that for every $\xi > 0$:

$$f(\widehat{x}_\mu) - f^* = \mathcal{O}\left[\left(\sqrt{\frac{\log(1/\delta)}{T}} + \upsilon + \xi\right)^{\beta_1 \beta_2}\right]$$

This combined with the fact that $f(\widehat{x}_{\widehat{\mu}}) \leq f(\widehat{x}_\mu)$ for every $\mu \in [-R, R]$ completes the proof of the main result (Thm. 2) .