Proceed ngs of Mach ne Learn ng esearch vo

Empirical Risk Minimization for Stochastic Convex Optimization: $O(1/n)$- and $O(1/n^2)$-type of Risk Bounds

Lijun Zhang ZHANGLJ LAMDA_NJ _ED _CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Tianbao Yang IANBAO YANG IO A_ED
Department of Computer Science, the University of Iowa, Iowa City, IA 52242, USA

Rong Jin JIN ONG_J ALIBABA INC_COM
Alibaba Group, Seattle, USA

Abstract

A though there ex st p ent fu theor es of e p r ca r s n zat on E M for superv sed ear n ng current theoret ca understand ngs of E M for a re ated prob e stochast c convex opt zat on CO are ted_ In th s wor we strengthen the rea of E M for CO by exp o t ng s oothness and strong convex ty cond t ons to prove the r s bounds_ F rst we estab sh an $\widetilde{\mathbf{O}}(\mathbf{d/n} + \sqrt{\mathbf{F_*/n}})$

where = · **h** : $\mathcal{X}$ $\mathbb{R}$ s a hypothes s c ass (**x**, **y**) $\mathcal{X}$ × $\mathbb{R}$ s an nstance abe pa r sa p ed fro a d str but on $\mathbb{D}$ and ℓ(·, ·) : $\mathbb{R}$ × $\mathbb{R}$ $\mathbb{R}$ s certa n oss_ In th s paper we a n y focus on the convex vers on of na e y stochast c convex opt zat on CO where both the do a n / and the expected funct on **F**(·) are convex_

wo c ass ca approaches for so v ng stochast c opt zat on are stochast c approx at on A Kushner and Y n ●● and the sa p e average approx at on AA the atter of wh ch s a so re ferred to as e p r ca r s n zat on E M n the ach ne earn ng co un ty apn 99 _ h e both A and E M have been extens ve y stud ed n recent years Bart ett and Mend e son ●● , Bart ett et a _ ●● , Ne rovs et a _ ●● 9, Mou nes and Bach ● , Hazan and Ka e ● , a h n et a _ ● , Agarwa et a _ ● , Bach and Mou nes ● , Zhang et a _ ● b, Mahdav et a _ ● ost theoret ca guarantees of E M are restr cted to superv sed earn ng n _As po nted out n a se na wor of ha ev hwartz et a _ ●● 9 the success of E M for su perv sed earn ng cannot be d rect y extended to stochast c opt zat on_Actua y ha ev hwartz et a _ ●● 9 have constructed an nstance of CO that s earnab e by A but cannot be so ved by E M_L teratures about E M for stochast c opt zat on nc ud ng CO are qu te ted and we st ac a fu understand ng of the theory_

In E M we are g ven **n** funct ons $\mathbf{f}_1, \ldots, \mathbf{f}_n$ sa p ed ndependent y fro $\mathbb{P}$ and a to n ze an e p r ca ob ect ve funct on·

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{\mathbf{F}}(\mathbf{w}) = \frac{1}{\mathbf{n}} \sum_{i=1}^{n} \mathbf{f}_i(\mathbf{w}).$$

Let $\widehat{\mathbf{w}}$ $\operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{\mathbf{F}}(\mathbf{w})$ be an e p r ca n zer_ he perfor ance of E M s easured n ter s of the excess r s de ned as

$$\mathbf{F}(\widehat{\mathbf{w}}) \quad \min_{\mathbf{w}\in\mathcal{W}} \mathbf{F}(\mathbf{w}).$$

tate of the art r s bounds of E M nc ude· an $\widetilde{\mathbf{O}}(\sqrt{\mathbf{d}/\mathbf{n}})$ bound when the rando funct on **f**(·) s L psch tz cont nuous where **d** s the d ens ona ty of **w**, an $\mathbf{O}(1/\boldsymbol{\lambda}\mathbf{n})$ bound when **f**(·) s **λ** strong y convex ha ev hwartz et a _ ●● 9 , and an $\widetilde{\mathbf{O}}(\mathbf{d}/\boldsymbol{\eta}\mathbf{n})$ bound when **f**(·) s **η** exponent a y concave **η** exp concave Mehta ● _ Fro ex st ng stud es of E M for superv sed earn ng rebro et a _ ● ● we now that s oothness can be ut zed to boost the r s bound_ hus t s natura to as whether s oothness can a so be exp o ted to prove the perfor ance of E M for CO_ h s paper prov des an af r at ve answer to th s quest on_Indeed we propose a genera approach for ana yz ng the excess r s of E M wh ch br ngs severa proved r s bounds and new r s bounds as we _

o state our resu ts we rst ntroduce so e notat ons_Let $\mathbf{F}_* = \min_{\mathbf{w}\in\mathcal{W}} \mathbf{F}(\mathbf{w})$ be the n a r s **λ** be the odu us of strong convex ty of **F**(·) and **L** be the odu us of s oothness of **f**(·)_ Denote by $\boldsymbol{\kappa} = \mathbf{L}/\boldsymbol{\lambda}$ the cond t on nu ber of the prob e _ Our and prev ous resu ts of E M for CO are su ar zed n ab e where we a e exp c t the assu pt ons on the rando funct on **f**(·) the e p r ca funct on $\widehat{\mathbf{F}}(\mathbf{w})$ and the expected funct on **F**(·)_ For our resu ts of E M for CO we assu e the do a n s bounded and the rando funct on s nonnegat ve_ e h gh ght the s gn cance of th s wor as fo ows·

_ e use the $\widetilde{O}$ and $\widetilde{\Omega}$ notat ons to h de constant factors as we as po y ogar th c factors n d and n_

ab e · u ary of Excess s Bounds of E M for CO_A bounds ho d w th h gh probab ty except the one ar ed by * wh ch ho ds n expectat on_ Abbrev at ons· bounded b convex c genera zed near g L psch tz cont nuous L p nonnegat ve nn strong y convex sc s ooth s $\boldsymbol{\eta}$ exponent a y concave $\boldsymbol{\eta}$ exp_

		$\mathbf{f}(\cdot)$	$\widehat{\mathbf{F}}(\cdot)$ $\mathbf{F}(\cdot)$	s Bounds
ha ev hwartz et a _ 9		L p		$\widetilde{\mathbf{O}}(\sqrt{\frac{d}{n}})$
		L p sc		$\mathbf{O}(\frac{1}{\lambda n})^*$
Mehta		$\boldsymbol{\eta}$ exp L p b		$\widetilde{\mathbf{O}}(\frac{d}{\eta n})$
h s wor	heore	nn c s	L p	$\widetilde{\mathbf{O}}(\frac{d}{n} + \sqrt{\frac{F_*}{n}})$
	heore	nn c s	L p sc	$\widetilde{\mathbf{O}}(\frac{d}{n} + \frac{\kappa F_*}{n})$ $\mathbf{O}(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n})$ when $\mathbf{n} = \widetilde{\Omega}(\boldsymbol{\kappa}\mathbf{d})$
	heore	nn s c sc		$\widetilde{\mathbf{O}}(\frac{\kappa d}{n} + \frac{\kappa F_*}{n}) = \widetilde{\mathbf{O}}(\frac{\kappa d}{n})$ $\mathbf{O}(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n})$ when $\mathbf{n} = \widetilde{\Omega}(\boldsymbol{\kappa}^2\mathbf{d})$
	heore	nn s g c sc		$\mathbf{O}(\frac{\kappa}{n} + \frac{\kappa F_*}{n}) = \mathbf{O}(\frac{\kappa}{n})$ $\mathbf{O}(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n})$ when $\mathbf{n} = \Omega(\boldsymbol{\kappa}^2)$

hen $\mathbf{f}(\cdot)$ s both convex and s ooth and $\mathbf{F}(\cdot)$ s L psch tz cont nuous we estab sh an $\widetilde{\mathbf{O}}(\mathbf{d}/\mathbf{n} + \sqrt{\mathbf{F}_*/\mathbf{n}})$ r s bound c f_ heore _ In the opt st c case that $\mathbf{F}_*$ s s a e_ $\mathbf{F}_* = \mathbf{O}(\mathbf{d}^2/\mathbf{n})$ we obta n an $\widetilde{\mathbf{O}}(\mathbf{d}/\mathbf{n})$ r s bound wh ch s ana ogous to the $\widetilde{\mathbf{O}}(1/\mathbf{n})$ opt st c rate of E M for superv sed earn ng rebro et a _ _

If $\mathbf{F}(\cdot)$ s a so $\boldsymbol{\lambda}$ strong y convex we prove an $\widetilde{\mathbf{O}}(\mathbf{d}/\mathbf{n} + \boldsymbol{\kappa}\mathbf{F}_*/\mathbf{n})$ r s bound and prove t to $\mathbf{O}(1/[\boldsymbol{\lambda}\mathbf{n}^2] + \boldsymbol{\kappa}\mathbf{F}_*/\mathbf{n})$ when $\mathbf{n} = \widetilde{\Omega}(\boldsymbol{\kappa}\mathbf{d})$ c f_ heore _ hus f $\mathbf{n}$ s arge and $\mathbf{F}_*$ s s a e_ $\mathbf{F}_* = \mathbf{O}(1/\mathbf{n})$ we get an $\mathbf{O}(\boldsymbol{\kappa}/\mathbf{n}^2)$ r s bound wh ch to the best of our now edge s the rst $\mathbf{O}(1/\mathbf{n}^2)$ type of r s bound of E M_

hen convex ty s not present n $\mathbf{f}(\cdot)$ as ong as $\mathbf{f}(\cdot)$ s s ooth $\widehat{\mathbf{F}}(\cdot)$ s convex and $\mathbf{F}(\cdot)$ s strong y convex we st obta n an proved r s bound of $\mathbf{O}(1/[\boldsymbol{\lambda}\mathbf{n}^2] + \boldsymbol{\kappa}\mathbf{F}_*/\mathbf{n})$ when $\mathbf{n} = \widetilde{\Omega}(\boldsymbol{\kappa}^2\mathbf{d})$ wh ch w further p y an $\mathbf{O}(\boldsymbol{\kappa}/\mathbf{n}^2)$ r s bound f $\mathbf{F}_* = \mathbf{O}(1/\mathbf{n})$ c f_ heore _ F na y we extend the $\mathbf{O}(1/[\boldsymbol{\lambda}\mathbf{n}^2] + \boldsymbol{\kappa}\mathbf{F}_*/\mathbf{n})$ r s bound to superv sed earn ng w th a genera zed near for _Our ana ys s shows that n th s case the ower bound of $\mathbf{n}$ can be rep aced w th $\Omega(\boldsymbol{\kappa}^2)$ wh ch s d ens ona ty ndependent c f_ heore _ hus th s resu t can be app ed to n n te d ens ona cases e g_ earn ng w th erne s_

2. Related Work

In th s sect on we g ve a br ef ntroduct on to prev ous wor on E M_

2.1. ERM for Stochastic Optimization

As we ent oned ear er there are few wor s devoted to E M for stochast c opt zat on_ hen / $\mathbb{R}^d$ s bounded and $\mathbf{f}(\cdot)$ s L psch tz cont nuous ha ev hwartz et a _ 09 de onstrate that $\widehat{\mathbf{F}}(\mathbf{w})$ converges to $\mathbf{F}(\mathbf{w})$ un for y over / w th an $\widetilde{\mathbf{O}}(\sqrt{\mathbf{d}/\mathbf{n}})$ error bound that ho ds w th h gh probab ty p y ng an $\widetilde{\mathbf{O}}(\sqrt{\mathbf{d}/\mathbf{n}})$ r s bound of E M_ hey further estab sh an $\mathbf{O}(1/\boldsymbol{\lambda}\mathbf{n})$ r s bound of E M that ho ds n expectat on when $\mathbf{f}(\cdot)$ s $\boldsymbol{\lambda}$ strong y convex and L psch tz con t nuous_ tochast c opt zat on w th exp concave funct ons s stud ed recent y Koren and Levy 0 and Mehta 0 proves an $\widetilde{\mathbf{O}}(\mathbf{d}/\boldsymbol{\eta}\mathbf{n})$ bound of E M that ho ds w th h gh probab ty when $\mathbf{f}(\cdot)$ s $\boldsymbol{\eta}$ exp concave L psch tz cont nuous and bounded_ Lower bounds of E M for stochast c opt zat on s nvest gated by Fe d an 0 who exh b ts a ower bound of $\Omega(\mathbf{d}/\square^2)$ sa p e co p ex ty for un for convergence that near y atches the upper bound of ha ev hwartz et a _ 09 , and a ower bound of $\Omega(\mathbf{d}/\square)$ sa p e co p ex ty of E M wh ch s atched by our $\widetilde{\mathbf{O}}(\mathbf{d}/\mathbf{n} + \sqrt{\mathbf{F}_*/\mathbf{n}})$ bound when $\mathbf{F}_*$ s s a _

2.2. ERM for Supervised Learning

e note that there are extens ve stud es on E M for superv sed earn ng and hence the rev ew here s non exhaust ve_ In the context of superv sed earn ng the perfor ance of E M s c ose y re ated to the un for convergence of $\widehat{\mathbf{F}}(\cdot)$ to $\mathbf{F}(\cdot)$ over the hypothes s c ass Ko tch ns 0 _ In fact un for convergence s a suf c ent cond t on for earnab ty ha ev hwartz and Ben Dav d 0 4 and n so e spec a cases such as b nary c ass cat on t s a so a necessary cond t on apn 99 _ he accuracy of un for convergence as we as the qua ty of the e p r ca n zer can be upper bounded n ter s of the co p ex ty of the hypothes s c ass nc ud ng data ndependent easures such as the C d ens on and data dependent easures such as the ade acher co p ex ty_

Genera y spea ng when has n te C d ens on the excess r s can be upper bounded by $\mathbf{O}(\sqrt{\mathrm{VC}(\ )/\mathbf{n}})$ where VC( ) s the C d ens on of _ If the oss $\square(\cdot,\cdot)$ s L psch tz con t nuous w th respect to ts rst argu ent we have a r s bound of $\mathbf{O}(1/\ \overline{\mathbf{n}} + \ _n(\ ))$ where $_n(\ )$ s the ade acher co p ex ty of _ he ade acher co p ex ty typ ca y sca es as $_n(\ ) = \mathbf{O}(1/\ \overline{\mathbf{n}})$ e g_ conta ns near funct ons w th ow nor p y ng an $\mathbf{O}(1/\ \overline{\mathbf{n}})$ r s bound Bart ett and Mende son 0 _ here have been ntens ve efforts to der ve rates faster than $\mathbf{O}(1/\ \overline{\mathbf{n}})$ under var ous cond t ons Lee et a _ 99 , Panchen o 0 , Bart ett et a _ 0 , Gonen and ha ev hwartz 0 such as ow no se syba ov 04 s oothness rebro et a _ 0 strong convex ty r dharan et a _ 09 to na e a few a ongst any_ pec ca y when the rando funct on $\mathbf{f}(\cdot)$ s nonnegat ve and s ooth rebro et a _ 0 have estab shed a r s bound of $\widetilde{\mathbf{O}}(\ ^2_n(\mathbf{H}) + \ _n(\mathbf{H})\ \mathbf{F}_*)$ reduc ng to an $\widetilde{\mathbf{O}}(1/\mathbf{n})$ bound f $_n(\ ) = \mathbf{O}(1/\ \overline{\mathbf{n}})$ and $\mathbf{F}_* = \mathbf{O}(1/\mathbf{n})$_ A genera zed near for of s stud ed by r dharan et a _ 09 and a r s bound of $\mathbf{O}(1/\boldsymbol{\lambda}\mathbf{n})$ s proved f the expected funct on $\mathbf{F}(\cdot)$ s $\boldsymbol{\lambda}$ strong y convex_

3. Faster Rates of ERM

e rst ntroduce a the assu pt ons used n our ana ys s then present theoret ca resu ts under d fferent co b nat ons of the and na y d scuss a spec a case of superv sed earn ng_

_ he r excess r s bound s for a regu ar zed e p r ca r s n zer_

3.1. Assumptions

In the fo ow ng we use $\| \cdot \|$ to denote the ℓ_2 nor of vectors_

Assumption 1 *The domain $\mathcal{W}$ is a convex subset of $\mathbb{R}^d$, and is bounded by R, that is,*

$$\|w\| \le R, \forall w \in \mathcal{W}. \tag{4}$$

Assumption 2 *The random function $f(\cdot)$ is nonnegative, and L-smooth over $\mathcal{W}$, that is,*

$$\|\nabla f(w) - \nabla f(w')\| \le L\|w - w'\|, \forall w, w' \in \mathcal{W}, f \in \mathbb{P}.$$

Assumption 3 *The expected function $F(\cdot)$ is G-Lipschitz continuous over $\mathcal{W}$, that is,*

$$|F(w) - F(w')| \le G\|w - w'\|, \forall w, w' \in \mathcal{W}.$$

Assumption 4 *We use different combinations of the following assumptions on convexity.*

(a) *The expected function $F(\cdot)$ is convex over $\mathcal{W}$.*

(b) *The expected function $F(\cdot)$ is λ-strongly convex over $\mathcal{W}$, that is,*

$$F(w) + \langle \nabla F(w), w' - w \rangle + \frac{\lambda}{2}\|w' - w\|^2 \le F(w'), \forall w, w' \in \mathcal{W}.$$

(c) *The empirical function $\widehat{F}(\cdot)$ is convex.*

(d) *The random function $f(\cdot)$ is convex.*

Assumption 5 *Let $w_* = \operatorname{argmin}_{w \in \mathcal{W}} F(w)$ be an optimal solution to (1). We assume the gradient of the random function at w_* is upper bounded by M, that is,*

$$\|\nabla f(w_*)\| \le M, \forall f \in \mathbb{P}.$$

Remark 1 F rst note that **Assumption 4(a)** s p ed by e ther **Assumption 4(b)** or **Assumption 4(d)** and **Assumption 4(c)** s p ed by **Assumption 4(d)**_ econd the s oothness assu p t on of $f(\cdot)$ p es the expected funct on $F(\cdot)$ s L s ooth_By Jensen s nequa ty we have

$$\|\nabla F(w) - \nabla F(w')\| = \|\mathbb{E}_{f \sim \mathbb{P}}[\nabla f(w) - \nabla f(w')]\| \le L\|w - w'\|, \forall w, w' \in \mathcal{W}.$$

ar y the e p r ca funct on $\widehat{F}(\cdot)$ s a so L s ooth_ he *condition number* κ of $F(\cdot)$ s de ned as the rat o between L and λ _e_ $\kappa = L/\lambda$ 1_

3.2. Risk Bounds for SCO

e rst present an excess r s bound under the s oothness cond t on_

Theorem 1 *For any $0 < \delta < 1/2$, $\varepsilon > 0$, define*

$$C(\varepsilon, \delta) = 2\left(\log\frac{2}{\delta} + d\log\frac{6R}{\varepsilon}\right). \tag{9}$$

Under **Assumptions** 1, 2, 3, 4(d)*, and* 5*, with probability at least* $1 \quad 2\delta$*, we have*

$$
\begin{aligned}
&\mathbf{F}(\widehat{\mathrm{w}}) \quad \mathbf{F}(\mathrm{w}_*) \\
&\frac{16\mathbf{R}^2\mathbf{L}\mathbf{C}(\epsilon,\delta)}{\mathbf{n}} + \frac{8\mathbf{R}\mathbf{M}\log(2/\delta)}{\mathbf{n}} + 8\mathbf{R}\sqrt{\frac{2\mathbf{L}\mathbf{F}_*\log(2/\delta)}{\mathbf{n}}} + \left(8\mathbf{R}\mathbf{L} + \mathbf{G} + \frac{4\mathbf{R}\mathbf{L}\mathbf{C}(\epsilon,\delta)}{\mathbf{n}}\right)\epsilon,
\end{aligned}
$$

where $\mathbf{F}_* = \mathbf{F}(\mathrm{w}_*)$ *is the minimal risk.*

By choos ng ϵ s a enough the ast ter n that conta ns ϵ beco es non do nat ng_ o be spec c we have the fo ow ng coro ary_

Corollary 2 *By setting* $\epsilon = 1/\mathbf{n}$ *in Theorem 1, we have* $\mathbf{C}(1/\mathbf{n},\delta) = 2(\log(2/\delta) + \mathbf{d}\log(6\mathbf{n}\mathbf{R}))$*, and with high probability*

$$
\mathbf{F}(\widehat{\mathrm{w}}) \quad \mathbf{F}(\mathrm{w}_*) = \mathbf{O}\left(\frac{\mathbf{d}\log\mathbf{n}}{\mathbf{n}} + \sqrt{\frac{\mathbf{F}_*}{\mathbf{n}}}\right) = \widetilde{\mathbf{O}}\left(\frac{\mathbf{d}}{\mathbf{n}} + \sqrt{\frac{\mathbf{F}_*}{\mathbf{n}}}\right).
$$

Remark 2 he above coro ary p es that under the s oothness and other co on assu pt ons E M ach eves an $\widetilde{\mathbf{O}}(\mathbf{d}/\mathbf{n} + \sqrt{\mathbf{F}_*/\mathbf{n}})$ r s bound for CO_ hen the n a r s s s a e_ $\mathbf{F}_* = \mathbf{O}(\mathbf{d}^2/\mathbf{n})$ the rate s proved to $\widetilde{\mathbf{O}}(\mathbf{d}/\mathbf{n})$_Note that even under the s oothness assu pt on the near dependence on $\mathbf{d}$ s unavo dab e Fe d an heore _ _

e next present excess r s bounds under both the s oothness and strong convex ty cond t ons_

Theorem 3 *Under* **Assumptions** 1, 2, 3, 4(b), 4(d)*, and* 5*, with probability at least* $1 \quad 2\delta$*, we have*

$$
\begin{aligned}
&\mathbf{F}(\widehat{\mathrm{w}}) \quad \mathbf{F}(\mathrm{w}_*) \\
&\frac{16\mathbf{R}^2\mathbf{L}\mathbf{C}(\epsilon,\delta)}{\mathbf{n}} + \frac{8\mathbf{R}\mathbf{M}\log(2/\delta)}{\mathbf{n}} + \frac{8\mathbf{L}\mathbf{F}_*\log(2/\delta)}{\lambda\mathbf{n}} + \left(8\mathbf{R}\mathbf{L} + \mathbf{G} + \frac{4\mathbf{R}\mathbf{L}\mathbf{C}(\epsilon,\delta)}{\mathbf{n}}\right)\epsilon.
\end{aligned}
$$

Furthermore, if

$$
\mathbf{n} \quad \frac{4\mathbf{L}\mathbf{C}(\epsilon,\delta)}{\lambda} = 4\kappa\mathbf{C}(\epsilon,\delta),
$$

we also have

$$
\mathbf{F}(\widehat{\mathrm{w}}) \quad \mathbf{F}(\mathrm{w}_*) \quad \frac{32\mathbf{M}^2\log^2(2/\delta)}{\lambda\mathbf{n}^2} + \frac{128\mathbf{L}\mathbf{F}_*\log(2/\delta)}{\lambda\mathbf{n}} + \left(\frac{128\mathbf{L}^2\epsilon^2}{\lambda} + 16\mathbf{G}\epsilon + 4\lambda\epsilon^2\right).
$$

he above theore can be s p ed by choos ng d fferent va ues of ϵ_

Corollary 4 *By setting* $\epsilon = 1/\mathbf{n}$ *in Theorem 3, we have* $\mathbf{C}(1/\mathbf{n},\delta) = 2(\log(2/\delta) + \mathbf{d}\log(6\mathbf{n}\mathbf{R}))$*, and with high probability*

$$
\mathbf{F}(\widehat{\mathrm{w}}) \quad \mathbf{F}(\mathrm{w}_*) = \mathbf{O}\left(\frac{\mathbf{d}\log\mathbf{n}}{\mathbf{n}} + \frac{\kappa\mathbf{F}_*}{\mathbf{n}}\right) = \widetilde{\mathbf{O}}\left(\frac{\mathbf{d}}{\mathbf{n}} + \frac{\kappa\mathbf{F}_*}{\mathbf{n}}\right).
$$

By setting $\epsilon = 1/\mathbf{n}^2$*, we have* $\mathbf{C}(1/\mathbf{n}^2,\delta) = 2\left(\log(2/\delta) + \mathbf{d}\log(6\mathbf{n}^2\mathbf{R})\right)$ *and when* $\mathbf{n} = \Omega(\kappa\mathbf{d}\log\mathbf{n}) = \widetilde{\Omega}(\kappa\mathbf{d})$*, with high probability*

$$
\mathbf{F}(\widehat{\mathrm{w}}) \quad \mathbf{F}(\mathrm{w}_*) = \mathbf{O}\left(\frac{1}{\lambda\mathbf{n}^2} + \frac{\kappa\mathbf{F}_*}{\mathbf{n}}\right).
$$

Remark 3 he rst part of Coro ary 4 shows that E M en oys an $\widetilde{O}(d/n + \kappa F_*/n)$ r s bound for stochast c opt zat on of strong y convex and s ooth funct ons_In the terature the ost co parab e resu t s the $O(1/\lambda n)$ r s bound proved by ha ev hwartz et a _ 09 but w th str ng d fferences h gh ghted n ab e _ nce the r s bound of ha ev hwartz et a _ 09 s ndepen dent of the d ens ona ty

Remark 6 Co par ng the second part of Coro ar es and 4 we can see that the r s bound s on the sa e order but the ower bound of $\mathbf{n}$ s ncreased by a factor of $\boldsymbol{\kappa}$_ It s nterest ng to ent on that a s ar pheno enon a so happens n stochast c approx at on_ ecent y a var ance reduct on techn que na ed G Johnson and Zhang or EMGD Zhang et a _ a was proposed for stochast c opt zat on when both fu grad ents and stochast c grad ents are ava ab e_ In the ana ys s G assu es the stochast c funct on s convex wh e EMGD does not_Fro the r theoret ca resu ts we observe that the nd v dua convex ty eads to a d fference of $\boldsymbol{\kappa}$ factor n the sa p e co p ex ty of stochast c grad ents_

3.3. Risk Bounds for Supervised Learning

If the cond t ons of heore or heore are sat s ed we can d rect y use the to estab sh an $\mathbf{O}(1/[\boldsymbol{\lambda}\mathbf{n}^2] + \boldsymbol{\kappa}\mathbf{F}_*/\mathbf{n})$ r s bound for superv sed earn ng_ However a a or tat on of these theore s s that the ower bound of $\mathbf{n}$ depends on the d ens ona ty $\mathbf{d}$ and thus cannot be app ed to n n te d ens ona cases e g_ erne ethods cho opf and o a _In th s sect on we exp o t the structure of superv sed earn ng to a e the theory d ens ona ty ndependent_

e focus on the genera zed near for of superv sed earn ng·

$$\min_{\mathbf{w}\in\mathcal{W}} \mathbf{F}(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\mathbf{w},\mathbf{x},\mathbf{y})\right] + \mathbf{r}(\mathbf{w}),$$

where $\ell(\mathbf{w},\mathbf{x},\mathbf{y})$ s the oss of pred ct ng $\mathbf{w},\mathbf{x}$ when the true target s $\mathbf{y}$ and $\mathbf{r}(\cdot)$ s a regu ar zer_G ven $\mathbf{n}$ tra n ng exa p es $(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)$ ndependent y sa p ed fro $\mathbb{D}$ the e p r ca ob ect ve s

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{\mathbf{F}}(\mathbf{w}) = \frac{1}{\mathbf{n}}\sum_{i=1}^{n}\ell(\mathbf{w},\mathbf{x}_i,\mathbf{y}_i) + \mathbf{r}(\mathbf{w}).$$

e de ne

$$\mathbf{H}(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\mathbf{w},\mathbf{x},\mathbf{y})\right] \text{ and } \widehat{\mathbf{H}}(\mathbf{w}) = \frac{1}{\mathbf{n}}\sum_{i=1}^{n}\ell(\mathbf{w},\mathbf{x}_i,\mathbf{y}_i)$$

to capture the stochast c co ponent_

Bes des 4**(b)** and 4**(c)** we ntroduce the fo ow ng add t ona assu pt ons_ e abuse the sa e notat on to denote the nor nduced by the nner product of a H bert space_

Assumption 6 *The domain is a convex subset of a Hilbert space , and is bounded by*

Assumption 10 *Let* $w_* = \operatorname{argmin}_{w \in \mathcal{W}} F(w)$ *be an optimal solution to (17). We assume the gradient of the random function at* w_* *is upper bounded by* M*, that is,*

$$\|\nabla \ell(\langle w_*, x\rangle, y)\| \leq M, \ \forall (x, y) \in \mathcal{D}.$$

Remark 7 The above assumptions allow us to model many popular losses in machine learning, such as regularized square loss and regularized logistic loss. Assumptions 7 and 8 imply the random function $\ell(\langle \cdot, x\rangle, y)$ is βD^2-smooth over $\mathcal{W}$. To see this, for any $w, w' \in \mathcal{W}$, we have

$$\begin{aligned} \|\nabla \ell(\langle w, x\rangle, y) - \nabla \ell(\langle w', x\rangle, y)\| &= \|\ell'(\langle w, x\rangle, y)x - \ell'(\langle w', x\rangle, y)x\| \\ &\overset{(9)}{\leq} D|\ell'(\langle w, x\rangle, y) - \ell'(\langle w', x\rangle, y)| \overset{(8)}{\leq} \beta D|\langle w, x\rangle - \langle w', x\rangle| \overset{(9)}{\leq} \beta D^2\|w - w'\|. \end{aligned}$$

By Jensen's inequality, $H(\cdot)$ is also βD^2-smooth. Notice that βD^2 is the modulus of smoothness of $H(\cdot)$ and λ is the modulus of strong convexity of $F(\cdot)$. With a slight abuse of notation, we define $L = \beta D^2$ and the condition number κ as the ratio between L and λ, i.e., $\kappa = L/\lambda$. Finally, we note that the regularizer $r(\cdot)$ could be *non-smooth*.

We have the following excess risk bound of ERM for supervised learning.

Theorem 7 *For any* $0 < \delta < 1/2$*, define*

$$C = 4\left(8 + \sqrt{2 \log \frac{2\log_2(n) + \log_2(2R)}{\delta}}\right),$$

$$H_* = H(w_*) = F(w_*) - r(w_*).$$

Under **Assumptions 4(b), 4(c), 6, 7, 8, 9,** *and* **10** *with probability at least* $1 - 2\delta$*, we have*

$$F(\widehat{w}) - F(w_*) \leq \max\left(\frac{M + P}{n^2} + \frac{L}{2n^4}, \frac{4R^2L^2C^2}{\lambda n} + \frac{4RM\log(2/\delta)}{n} + \frac{8LH_*\log(2/\delta)}{\lambda n}\right).$$

Furthermore, if

$$n \geq \frac{16L^2C^2}{\lambda^2} = 16\kappa^2C^2,$$

with probability at least $1 - 2\delta$*, we have*

$$F(\widehat{w}) - F(w_*) \leq \max\left(\frac{M + P}{n^2} + \frac{L}{2n^4}, \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{16LH_*\log(2/\delta)}{\lambda n}\right).$$

Remark 8 The first part of Theorem 7 presents an $O(\kappa/n)$ risk bound, similar to the $O(1/\lambda n)$ risk bound of Sridharan et al., 2009. The second part is an $O(1/[\lambda n^2] + \kappa H_*/n)$ risk bound, and in this case, the lower bound of n is $\Omega(\kappa^2)$, which is dimensionality-independent. Thus, Theorem 7 can be applied even when the dimensionality is infinite. Generally speaking, the regularizer $r(\cdot)$ is nonnegative, and thus $H_* \leq F_*$. So the second bound is even better than those in Theorems [illegible] and [illegible]. Finally, we note that Theorem 7 should be treated as a counterpart of Theorem [illegible] for supervised learning, because both of them do not rely on the individual convexity (i.e., Assumption 4(d)). One may wonder whether it is possible to derive a counterpart of Theorem [illegible], that is, whether it is possible to utilize the individual convexity to reduce the lower bound of n by a factor of κ. We will investigate this question as a future work.

For brevity, we treat C as a constant because it only has a *double* logarithmic dependence on n.

4. Analysis

e here present the ey dea of our ana ys s and the proof of heore _ he o tted ones can be found n append ces_

4.1. The Key Idea

By the convex ty of $\widehat{\mathbf{F}}(\cdot)$ and the opt a ty cond t on of $\widehat{\mathbf{w}}$ Boyd and andenberghe 004 we have

Lemma 1 *Under* **Assumptions 2** *and* **4(d)**, *with probability at least* 1 δ

where the ast step s due to

$$\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})(\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*))}{\mathbf{n}}} \quad \frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta}) \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2}{2\mathbf{n}} + \frac{\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*)}{2},$$

$$\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\mathbf{G}\boldsymbol{\varepsilon}}{\mathbf{n}}} \quad \frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta}) \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2}{2\mathbf{n}} + \frac{\mathbf{G}\boldsymbol{\varepsilon}}{2}.$$

Fro 4 we get

$$\begin{aligned}
&\frac{1}{2}(\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*)) \\
&\frac{2\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta}) \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2}{\mathbf{n}} + \frac{2\mathbf{M}\log(2/\boldsymbol{\delta}) \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|}{\mathbf{n}} + \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{8\mathbf{LF}_*\log(2/\boldsymbol{\delta})}{\mathbf{n}}} \\
&+ 2\mathbf{L}\boldsymbol{\varepsilon} \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| + \frac{\mathbf{G}\boldsymbol{\varepsilon}}{2} + \frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\boldsymbol{\varepsilon} \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|}{\mathbf{n}} \\
4 \quad &\frac{8\mathbf{R}^2\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}} + \frac{4\mathbf{RM}\log(2/\boldsymbol{\delta})}{\mathbf{n}} + 4\mathbf{R}\sqrt{\frac{2\mathbf{LF}_*\log(2/\boldsymbol{\delta})}{\mathbf{n}}} + \left(4\mathbf{RL} + \frac{\mathbf{G}}{2} + \frac{2\mathbf{RLC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}\right)\boldsymbol{\varepsilon},
\end{aligned}$$

wh ch p es

5. Conclusions and Future Work

In th s paper we study the excess r s of E M for CO_Our theoret ca resu ts show that t s poss b e to ach eve $\mathbf{O}(1/\mathbf{n})$ type of r s bounds under the s oothness and s a n a r s cond t ons _e_ heore or the s oothness and strong convex ty cond t ons _e_ the rst part of heore s and _A ore exc t ng resu t s that when $\mathbf{n}$ s arge enough E M has $\mathbf{O}(1/\mathbf{n}^2)$ type of r s bounds under the s oothness strong convex ty and s a n a r s cond t ons _e_ the second part of heore s and _

In the context of CO there re a n any open prob e s about E M_

- _ Our current resu ts are restr cted to the H bert or Euc dean space because the s oothness and strong convex ty are de ned n ter s of the ℓ_2 nor _ e w extend our ana ys s to other geo etr es n the future_
- _ As ent oned n **Remark 3** under the strong convex ty cond t on a d ens ona ty ndependent r s bound e g_ $\widetilde{\mathbf{O}}(\boldsymbol{\kappa}/\mathbf{n})$ or $\widetilde{\mathbf{O}}(1/\boldsymbol{\lambda}\mathbf{n})$ that ho ds w th h gh probab ty s st ss ng_
- _ As d scussed n **Remark 8** t s unc ear whether the convex ty of the oss can be exp o ted to prove the ower bound of $\mathbf{n}$ n the second part of heore _ Idea y we expect that $\mathbf{n} = \Omega(\boldsymbol{\kappa})$ s suf c ent to de ver an $\mathbf{O}(1/[\boldsymbol{\lambda}\mathbf{n}^2] + \boldsymbol{\kappa}\mathbf{H}_*/\mathbf{n})$ r s bound_
- 4_ he $\mathbf{O}(1/\mathbf{n}^2)$ type of r s bounds requ re both the s oothness and strong convex ty cond t ons_One ay nvest gate whether strong convex ty can be re axed to other wea er cond t ons such as exponent a concav ty Hazan et a _ _

F na y as far as we now there are no $\mathbf{O}(1/\mathbf{n}^2)$ type of r s bounds for stochast c approx at on A _ e w try to estab sh such bounds for A_

Acknowledgments

h s wor was part a y supported by the N FC J angsu F BK N F II 4 9 II 4 99 and the Co aborat ve Innovation Center of Nove oftware echno ogy and Industr a zat on of Nan ng n vers ty_

References

A e h Agarwa Peter L_Bart ett Pradeep av u ar and Mart n J_ a nwr ght_ Infor at on theoret c ower bounds on the orac e co p ex ty of stochast c convex opt zat on_ *IEEE Transactions on Information Theory* · 49 _

Franc s Bach and Er c Mou nes_Non strong y convex s ooth stochast c approx at on w th con vergence rate $O(1/n)$_ In *Advances in Neural Information Processing Systems 26* pages _

Peter L_Bart ett and hahar Mende son_ ade acher and gauss an co p ex t es· r s bounds and structura resu ts_ *Journal of Machine Learning Research* ·4 4 _

Peter L_Bart ett O v er Bousquet and hahar Mende son_ Loca rade acher co p ex t es_ *The Annals of Statistics* 4 · 49 _

tephen Boyd and L even andenberghe_ *Convex Optimization*_Ca br dge n vers ty Press 4_

G u a Desa vo Mehryar Mohr and ar yed_ Learn ng w th deep cascades_ In *Proceedings of the 26th International Conference on Algorithmic Learning Theory* pages 4 9 _

ta y Fe d an_ Genera zat on of er n stochast c convex opt zat on· he d ens on str es bac _ *ArXiv e-prints* arXiv v· 44 4 _

A on Gonen and ha ha ev hwartz_Average stab ty s nvar ant to data precond t on ng_ p cat ons to exp concave e p r ca r s n zat on_ *ArXiv e-prints* arXiv v· _

E ad Hazan and atyen Ka e_Beyond the regret n zat on barr er· an opt a a gor th for sto chast c strong y convex opt zat on_In *Proceedings of the 24th Annual Conference on Learning Theory* pages 4 4 _

E ad Hazan A t Agarwa and atyen Ka e_ Logar th c regret a gor th s for on ne convex opt zat on_ *Machine Learning* 9 · 9 9 _

e Johnson and ong Zhang_ Acce erat ng stochast c grad ent descent us ng pred ct ve var ance reduct on_In *Advances in Neural Information Processing Systems 26* pages _

ad r Ko tch ns _ *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*_ pr nger _

o er Koren and K r Levy_ Fast rates for exp concave e p r ca r s n zat on_ In *Advances in Neural Information Processing Systems 28* pages 4 4 _

Haro d J_Kushner and G_George Y n_ *Stochastic Approximation and Recursive Algorithms and Applications*

Mehrdad Mahdav L un Zhang and ong J n_ Lower and upper bounds on the genera zat on of stochast c exponent a y concave opt zat on_ In *Proceedings of the 28th Conference on Learning Theory* _

Co n McD ar d_ On the ethod of bounded d fferences_ In *Surveys in Combinatorics* pages 9 9_

N shant A_Mehta_ Fast rates w th h gh probab ty n exp concave stat st ca earn ng_ *ArXiv e-prints* arX v _

on Me r and ong Zhang_ Genera zat on error bounds for bayes an xture a gor th s_ *Journal of Machine Learning Research* 9 _

Er c Mou nes and Franc s _Bach_ Non asy ptot c ana ys s of stochast c approx at on a go r th s for ach ne earn ng_ In *Advances in Neural Information Processing Systems 24* pages 9 _

A_Ne rovs A_Jud ts y G_Lan and A_ hap ro_ obust stochast c approx at on approach to stochast c progra ng_ *SIAM Journal on Optimization* 9 9_

Yur Nesterov_ *Introductory lectures on convex optimization: a basic course* vo u e of *Applied optimization*_K uwer Acade c Pub shers _

D tr y Panchen o_ o e extens ons of an nequa ty of vapn and chervonen s_ *Electronic Communications in Probability* _

G es P s er_ *The volume of convex bodies and Banach space geometry*_ Ca br dge racts n Mathe at cs No_94 _Ca br dge n vers ty Press 9 9_

Yan v P an and o an ershyn n_One b t co pressed sens ng by near progra ng_ *Communications on Pure and Applied Mathematics* 9_ _

A exander a h n Ohad ha r and Karth r dharan_ Ma ng grad ent descent opt a for strong y convex stochast c opt zat on_In *Proceedings of the 29th International Conference on Machine Learning* pages 9 _

Bernhard chò opf and A exander J_ o a_ *Learning with kernels: support vector machines, regularization, optimization, and beyond*_MI Press _

ha ha ev hwartz and ha Ben Dav d_ *Understanding Machine Learning: From Theory to Algorithms*_Ca br dge n vers ty Press 9_

ha ha ev hwartz Ohad ha r Nathan rebro and Karth r dharan_ tochast c convex opt zat on_In *Proceedings of the 22nd Annual Conference on Learning Theory* 9_

A exander hap ro Dar n a Dentcheva and Andrze uszczyńs _ *Lectures on Stochastic Programming: Modeling and Theory*_ IAM second ed t on _

teve a e and D ng Xuan Zhou_ Learn ng theory est ates v a ntegra operators and the r ap prox at ons_ *Constructive Approximation* _

Nathan rebro Karth r dharan and A bu ewar _ Opt st c rates for earn ng w th a s ooth oss_*ArXiv e-prints* arX v· ●● 9_ 9 ● ●_

Karth r dharan ha ha ev shwartz and Nathan rebro_ Fast rates for regu ar zed ob ect ves_ In *Advances in Neural Information Processing Systems 21* pages 4 ●● 9_

A exandre B_ syba ov_ Opt a aggregat on of c ass ers n stat st ca earn ng_ *The Annals of Statistics* · ●● 4_

ad r apn _*The Nature of Statistical Learning Theory*_ pr nger second ed t on ●●● _

ad r N_ apn _*Statistical Learning Theory*_ ey Intersc ence 99 _

L un Zhang Mehrdad Mahdav and ong J n_ L near convergence w th cond t on nu ber nde pendent access of fu grad ents_In *Advance in Neural Information Processing Systems 26* pages 9● 9 ● a_

L un Zhang anbao Yang ong J n and X aofe He_ $\mathbf{O}(\log \mathbf{T})$ pro ect ons for stochast c op t zat on of s ooth and strong y convex funct ons_ In *Proceedings of the 30th International Conference on Machine Learning* ● b_

Appendix A. Proof of Lemma 1

e ntroduce Le a of a e and Zhou ●● _

Lemma 3 *Let be a Hilbert space and let* $\boldsymbol{\xi}$ *be a random variable with values in . Assume* $\boldsymbol{\xi}$ $\mathbf{M} <$ *almost surely. Denote* $\boldsymbol{\sigma}^2(\boldsymbol{\xi}) = \mathrm{E}\left[\boldsymbol{\xi}^2\right]$. *Let* $\cdot \boldsymbol{\xi}_i \,_{i=1}^{m}$ *be* $\mathbf{m}$ *(*$\mathbf{m} <$ *) independent drawers of* $\boldsymbol{\xi}$. *For any* $0 < \boldsymbol{\delta} < 1$, *with confidence* 1 $\boldsymbol{\delta}$,

$$\left\| \frac{1}{\mathbf{m}} \sum_{i=1}^{m} [\boldsymbol{\xi}_i \quad \mathrm{E}[\boldsymbol{\xi}_i]] \right\| \quad \frac{2\mathbf{M}\log(2/\boldsymbol{\delta})}{\mathbf{m}} + \sqrt{\frac{2\boldsymbol{\sigma}^2(\boldsymbol{\xi})\log(2/\boldsymbol{\delta})}{\mathbf{m}}}.$$

e rst cons der a xed $\mathbf{w}$ $(\mathbf{,} \boldsymbol{\varepsilon})$_ nce $\mathbf{f}_i(\,)$ s $\mathbf{L}$ s ooth we have

$$\mathbf{f}_i(\mathbf{w}) \quad \mathbf{f}_i(\mathbf{w}_*) \quad \mathbf{L} \, \mathbf{w} \quad \mathbf{w}_* .$$

Because $\mathbf{f}_i(\,)$ s both convex and $\mathbf{L}$ s ooth by of Nesterov ●● 4 we have

$$\mathbf{f}_i(\mathbf{w}) \quad \mathbf{f}_i(\mathbf{w}_*)^2 \quad \mathbf{L}\,(\mathbf{f}_i(\mathbf{w}) \quad \mathbf{f}_i(\mathbf{w}_*) \quad \mathbf{f}_i(\mathbf{w}_*), \mathbf{w} \quad \mathbf{w}_*).$$

a ng expectat on over both s des we have

$$\mathrm{E}\left[\mathbf{f}_i(\mathbf{w}) \quad \mathbf{f}_i(\mathbf{w}_*)^2\right] \quad \mathbf{L}\,(\mathbf{F}(\mathbf{w}) \quad \mathbf{F}(\mathbf{w}_*) \quad \mathbf{F}(\mathbf{w}_*), \mathbf{w} \quad \mathbf{w}_*) \quad \mathbf{L}\,(\mathbf{F}(\mathbf{w}) \quad \mathbf{F}(\mathbf{w}_*))$$

where the ast nequa ty fo ows fro the opt a ty cond t on of $\mathbf{w}_*$ _e_

$$\mathbf{F}(\mathbf{w}_*), \mathbf{w} \quad \mathbf{w}_* \quad 0, \bullet \mathbf{w} \quad \mathbf{,} .$$

Fo ow ng Le a w th probab ty at east 1 δ we have

$$\left\| \mathbf{F}(\mathbf{w}) \quad \mathbf{F}(\mathbf{w}_*) \quad [\widehat{\mathbf{F}}(\mathbf{w}) \quad \widehat{\mathbf{F}}(\mathbf{w}_*)] \right\| = \left\| \mathbf{F}(\mathbf{w}) \quad \mathbf{F}(\mathbf{w}_*) \quad \frac{1}{\mathbf{n}} \sum_{i=1}^{n} [\mathbf{f}_i(\mathbf{w}) \quad \mathbf{f}_i(\mathbf{w}_*)] \right\|$$
$$\frac{2\mathbf{L} \mathbf{w} \quad \mathbf{w}_* \log(2/\delta)}{\mathbf{n}} + \sqrt{\frac{2\mathbf{L}(\mathbf{F}(\mathbf{w}) \quad \mathbf{F}(\mathbf{w}_*)) \log(2/\delta)}{\mathbf{n}}}.$$

e obta n Le a by ta ng the un on bound over a w (, ε) o th s end we need an upper bound of the cover ng nu ber (, ε) .

Let $\mathcal{B}$ be an un t ba of d d ens on and $(\mathcal{B}, \varepsilon)$ be ts ε net w th n a card na ty. Accor d ng to a standard vo u e co par son argu ent P s er 9 9 we have

$$\log (\mathcal{B}, \varepsilon) \quad \mathbf{d} \log \frac{3}{\varepsilon}.$$

Let $\mathcal{B}(\mathbf{R})$ be a ba centered at or g n w th rad us R nce we assu e $\mathcal{B}(\mathbf{R})$ t fo ows that

$$\log (, \varepsilon) \quad \log \left| \left(\mathcal{B}(\mathbf{R}), \frac{\varepsilon}{2}\right) \right| \quad \mathbf{d} \log \frac{6\mathbf{R}}{\varepsilon}$$

where the rst nequa ty s because the cover ng nu bers are a ost ncreas ng by nc us on P an and ershyn n .

Appendix B. Proof of Lemma 2

o app y Le a we need an upper bound of $\mathrm{E}\left[\mathbf{f}_i(\mathbf{w}_*) ^2\right]$ nce $\mathbf{f}_i( )$ s L s ooth and nonnegat ve fro Le a 4 of rebro et a . we have

$$\mathbf{f}_i(\mathbf{w}_*) ^2 \quad 4\mathbf{L}\mathbf{f}_i(\mathbf{w}_*)$$

and thus

$$\mathrm{E}\left[\mathbf{f}_i(\mathbf{w}_*) ^2\right] \quad 4\mathbf{L}\mathrm{E}\left[\mathbf{f}_i(\mathbf{w}_*)\right] = 4\mathbf{L}\mathbf{F}_*.$$

Fro **Assumption 5** we have $\mathbf{f}_i(\mathbf{w}_*) \quad \mathbf{M}$ hen accord ng to Le a w th probab ty at east 1 δ we have

$$\left\| \mathbf{F}(\mathbf{w}_*) \quad \widehat{\mathbf{F}}(\mathbf{w}_*) \right\| = \left\| \mathbf{F}(\mathbf{w}_*) \quad \frac{1}{\mathbf{n}} \sum_{i=1}^{n} \mathbf{f}_i(\mathbf{w}_*) \right\| \quad \frac{2\mathbf{M} \log(2/\delta)}{\mathbf{n}} + \sqrt{\frac{8\mathbf{L}\mathbf{F}_* \log(2/\delta)}{\mathbf{n}}}.$$

Appendix C. Proof of Theorem 3

he proof fo ows the sa e og c as that of heore nder **Assumption 4(b)** beco es

$$\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*) + \frac{\lambda}{2} \widehat{\mathbf{w}} \quad \mathbf{w}_* ^2$$
$$\left(\underbrace{\left\| \mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*) \quad [\widehat{\mathbf{F}}(\widehat{\mathbf{w}}) \quad \widehat{\mathbf{F}}(\mathbf{w}_*)] \right\|}_{:=A_1} + \underbrace{\left\| \mathbf{F}(\mathbf{w}_*) \quad \widehat{\mathbf{F}}(\mathbf{w}_*) \right\|}_{:=A_2} \right) \widehat{\mathbf{w}} \quad \mathbf{w}_* .$$

ubst tut ng and nto w th probab ty at east $1 \quad 2\delta$ we have

$$\begin{aligned} F(\widehat{w}) \quad F(w_*) &+ \frac{\lambda}{2} \widehat{w} \quad w_* ^2 \\ &\frac{LC(\varepsilon,\delta)}{n} \widehat{w} \quad w_* ^2 + \widehat{w} \quad w_* \sqrt{\frac{LC(\varepsilon,\delta)(F(\widehat{w}) \quad F(w_*))}{n}} \\ &+ \frac{2M \log(2/\delta)}{n} \widehat{w} \quad w_* + \widehat{w} \quad w_* \sqrt{\frac{8LF_* \log(2/\delta)}{n}} \\ &+ 2L\varepsilon \widehat{w} \quad w_* + \widehat{w} \quad w_* \sqrt{\frac{LC(\varepsilon,\delta)G\varepsilon}{n}} + \frac{LC(\varepsilon,\delta)\varepsilon}{n} \widehat{w} \quad w_* \end{aligned} \tag{9}$$

o prove we subst tute and

$$\widehat{w} \quad w_* \sqrt{\frac{8LF_* \log(2/\delta)}{n}} \quad \frac{4LF_* \log(2/\delta)}{\lambda n} + \frac{\lambda}{2} \widehat{w} \quad w_* ^2$$

nto 9 and then obta n

$$\begin{aligned} &\frac{1}{2}(F(\widehat{w}) \quad F(w_*)) \\ &\quad \frac{2LC(\varepsilon,\delta)}{n} \widehat{w} \quad w_* ^2 + \frac{2M \log(2/\delta)}{n} \widehat{w} \quad w_* + \frac{4LF_* \log(2/\delta)}{\lambda n} \\ &+ 2L\varepsilon \widehat{w} \quad w_* + \frac{G\varepsilon}{2} + \frac{LC(\varepsilon,\delta)\varepsilon}{n} \widehat{w} \quad w_* \\ &\quad \frac{8R^2 LC(\varepsilon,\delta)}{n} + \frac{4RM \log(2/\delta)}{n} + \frac{4LF_* \log(2/\delta)}{\lambda n} + \left(4RL + \frac{G}{2} + \frac{2RLC(\varepsilon,\delta)}{n}\right)\varepsilon. \end{aligned}$$

wh ch p es

o prove we subst tute

$$\begin{aligned} \widehat{w} \quad w_* \sqrt{\frac{LC(\varepsilon,\delta)(F(\widehat{w}) \quad F(w_*))}{n}} \quad & \frac{2LC(\varepsilon,\delta)(F(\widehat{w}) \quad F(w_*))}{\lambda n} + \frac{\lambda}{8} \widehat{w} \quad w_* ^2, \\ \frac{2M \log(2/\delta)}{n} \widehat{w} \quad w_* \quad & \frac{16M^2 \log^2(2/\delta)}{\lambda n^2} + \frac{\lambda}{16} \widehat{w} \quad w_* ^2, \\ \widehat{w} \quad w_* \sqrt{\frac{8LF_* \log(2/\delta)}{n}} \quad & \frac{64LF_* \log(2/\delta)}{\lambda n} + \frac{\lambda}{32} \widehat{w} \quad w_* ^2, \\ 2L\varepsilon \widehat{w} \quad w_* \quad & \frac{64L^2\varepsilon^2}{\lambda} + \frac{\lambda}{64} \widehat{w} \quad w_* ^2, \\ \widehat{w} \quad w_* \sqrt{\frac{LC(\varepsilon,\delta)G\varepsilon}{n}} \quad & \frac{32LC(\varepsilon,\delta)G\varepsilon}{\lambda n} + \frac{\lambda}{128} \widehat{w} \quad w_* ^2, \\ \frac{LC(\varepsilon,\delta)\varepsilon}{n} \widehat{w} \quad w_* \quad & \frac{32L^2C^2(\varepsilon)\varepsilon^2}{\lambda n^2} + \frac{\lambda}{128} \widehat{w} \quad w_* ^2 \end{aligned}$$

nto 9 and then obta n

$$\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*) + \frac{\boldsymbol{\lambda}}{4}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2$$
$$\frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2}{\mathbf{n}} + \frac{2\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})(\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*))}{\boldsymbol{\lambda}\mathbf{n}} + \frac{16\mathbf{M}^2\log^2(2/\boldsymbol{\delta})}{\boldsymbol{\lambda}\mathbf{n}^2} + \frac{64\mathbf{LF}_*\log(2/\boldsymbol{\delta})}{\boldsymbol{\lambda}\mathbf{n}}$$
$$+ \frac{64\mathbf{L}^2\boldsymbol{\varepsilon}^2}{\boldsymbol{\lambda}} + \frac{32\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\mathbf{G}\boldsymbol{\varepsilon}}{\boldsymbol{\lambda}\mathbf{n}} + \frac{32\mathbf{L}^2\mathbf{C}^2(\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}^2}{\boldsymbol{\lambda}\mathbf{n}^2}$$
$$\frac{\boldsymbol{\lambda}}{4}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2 + \frac{1}{2}(\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*)) + \frac{16\mathbf{M}^2\log^2(2/\boldsymbol{\delta})}{\boldsymbol{\lambda}\mathbf{n}^2} + \frac{64\mathbf{LF}_*\log(2/\boldsymbol{\delta})}{\boldsymbol{\lambda}\mathbf{n}}$$
$$+ \frac{64\mathbf{L}^2\boldsymbol{\varepsilon}^2}{\boldsymbol{\lambda}} + 8\mathbf{G}\boldsymbol{\varepsilon} + 2\boldsymbol{\lambda}\boldsymbol{\varepsilon}^2$$

wh ch p es _

Appendix D. Proof of Theorem 5

thout **Assumption 4(d)** Le a wh ch s used n the proofs of heore s and does not ho d any ore_Instead we w use the fo ow ng vers on that on y re es on the s oothness cond t on_

Lemma 4 *Under* **Assumption 2**, *with probability at least* $1 \quad \boldsymbol{\delta}$, *for any* $\mathbf{w} \quad (\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})$, *we have*

$$\left\| \mathbf{F}(\mathbf{w}) \quad \mathbf{F}(\mathbf{w}_*) \quad [\widehat{\mathbf{F}}(\mathbf{w}) \quad \widehat{\mathbf{F}}(\mathbf{w}_*)] \right\| \quad \frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\left\| \mathbf{w} \quad \mathbf{w}_* \right\|}{\mathbf{n}} + \mathbf{L}\left\| \mathbf{w} \quad \mathbf{w}_* \right\| \sqrt{\frac{\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}}$$

where $\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})$ *is define in (9).*

he above e a s a d rect consequence of Le a and the un on bound_ he rest of the proof s s ar to those of heore s and _ e rst der ve a counterpart of under Le a 4_Co b n ng w th Le a 4 w th probab ty at east $1 \quad \boldsymbol{\delta}$ we have

$$\left\| \mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*) \quad [\widehat{\mathbf{F}}(\widehat{\mathbf{w}}) \quad \widehat{\mathbf{F}}(\mathbf{w}_*)] \right\|$$
$$\frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\left\| \widetilde{\mathbf{w}} \quad \mathbf{w}_* \right\|}{\mathbf{n}} + \mathbf{L}\left\| \widetilde{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}} + 2\mathbf{L}\boldsymbol{\varepsilon} \qquad (40)$$
$$\frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|}{\mathbf{n}} + \mathbf{L}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}} + \frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\boldsymbol{\varepsilon}}{\mathbf{n}} + \mathbf{L}\boldsymbol{\varepsilon}\sqrt{\frac{\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}} + 2\mathbf{L}\boldsymbol{\varepsilon}.$$

ubst tut ng 40 and nto w th probab ty at east $1 \quad 2\boldsymbol{\delta}$ we have

$$\mathbf{F}(\widehat{\mathbf{w}}) \quad \mathbf{F}(\mathbf{w}_*) + \frac{\boldsymbol{\lambda}}{2}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2$$
$$\frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2}{\mathbf{n}} + \mathbf{L}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|^2 \sqrt{\frac{\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}}$$
$$+ \frac{2\mathbf{M}\log(2/\boldsymbol{\delta})\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|}{\mathbf{n}} + \left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{8\mathbf{LF}_*\log(2/\boldsymbol{\delta})}{\mathbf{n}}} \qquad (41)$$
$$+ 2\mathbf{L}\boldsymbol{\varepsilon}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| + \mathbf{L}\boldsymbol{\varepsilon}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\| \sqrt{\frac{\mathbf{C}(\boldsymbol{\varepsilon},\boldsymbol{\delta})}{\mathbf{n}}} + \frac{\mathbf{LC}(\boldsymbol{\varepsilon},\boldsymbol{\delta})\boldsymbol{\varepsilon}\left\| \widehat{\mathbf{w}} \quad \mathbf{w}_* \right\|}{\mathbf{n}}.$$

o get 4 we subst tute

$$L\,|\,\widehat{w} \quad w_*\,|^2\sqrt{\frac{C(\varepsilon,\delta)}{n}} \quad \frac{L^2C(\varepsilon,\delta)}{\lambda n}\,|\,\widehat{w} \quad w_*\,|^2 + \frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$|\,\widehat{w} \quad w_*\,|\sqrt{\frac{8LF_*\log(2/\delta)}{n}} \quad \frac{8LF_*\log(2/\delta)}{\lambda n} + \frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2$$

nto 4 and then obta n

$$F(\widehat{w}) \quad F(w_*)$$
$$\frac{LC(\varepsilon,\delta)}{n}\,|\,\widehat{w} \quad w_*\,|^2 + \frac{L^2C(\varepsilon,\delta)}{\lambda n}\,|\,\widehat{w} \quad w_*\,|^2 + \frac{2M\log(2/\delta)}{n}\,|\,\widehat{w} \quad w_*\,| + \frac{8LF_*\log(2/\delta)}{\lambda n}$$
$$+ 2L\varepsilon\,|\,\widehat{w} \quad w_*\,| + L\varepsilon\,|\,\widehat{w} \quad w_*\,|\sqrt{\frac{C(\varepsilon,\delta)}{n}} + \frac{LC(\varepsilon,\delta)\varepsilon}{n}\,|\,\widehat{w} \quad w_*\,|$$
$$4\,\frac{4R^2LC(\varepsilon,\delta)}{n} + \frac{4R^2L^2C(\varepsilon,\delta)}{\lambda n} + \frac{4RM\log(2/\delta)}{n} + \frac{8LF_*\log(2/\delta)}{\lambda n}$$
$$+ \left(4RL + 2RL\sqrt{\frac{C(\varepsilon,\delta)}{n}} + \frac{2RLC(\varepsilon,\delta)}{n}\right)\varepsilon$$

wh ch proves 4 _
o get we subst tute

$$\frac{2M\log(2/\delta)}{n}\,|\,\widehat{w} \quad w_*\,| \quad \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{\lambda}{8}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$|\,\widehat{w} \quad w_*\,|\sqrt{\frac{8LF_*\log(2/\delta)}{n}} \quad \frac{32LF_*\log(2/\delta)}{\lambda n} + \frac{\lambda}{16}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$2L\varepsilon\,|\,\widehat{w} \quad w_*\,| \quad \frac{32L^2\varepsilon^2}{\lambda} + \frac{\lambda}{32}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$L\varepsilon\,|\,\widehat{w} \quad w_*\,|\sqrt{\frac{C(\varepsilon,\delta)}{n}} \quad \frac{16L^2C(\varepsilon,\delta)\varepsilon^2}{\lambda n} + \frac{\lambda}{64}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$\frac{LC(\varepsilon,\delta)\varepsilon}{n}\,|\,\widehat{w} \quad w_*\,| \quad \frac{16L^2C^2(\varepsilon)\varepsilon^2}{\lambda n^2} + \frac{\lambda}{64}\,|\,\widehat{w} \quad w_*\,|^2$$

nto 4 and then obta n

$$F(\widehat{w}) \quad F(w_*) + \frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2$$
$$\frac{LC(\varepsilon,\delta)}{n}\,|\,\widehat{w} \quad w_*\,|^2 + L\,|\,\widehat{w} \quad w_*\,|^2\sqrt{\frac{C(\varepsilon,\delta)}{n}} + \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{32LF_*\log(2/\delta)}{\lambda n}$$
$$+ \left(\frac{32L^2}{\lambda} + \frac{16L^2C(\varepsilon,\delta)}{\lambda n} + \frac{16L^2C^2(\varepsilon)}{\lambda n^2}\right)\varepsilon^2$$
$$\frac{\lambda^2}{25L}\,|\,\widehat{w} \quad w_*\,|^2 + \frac{\lambda}{5}\,|\,\widehat{w} \quad w_*\,|^2 + \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{32LF_*\log(2/\delta)}{\lambda n}$$
$$+ \left(\frac{32L^2}{\lambda} + \frac{16\lambda}{25} + \frac{16\lambda^3}{625L^2}\right)\varepsilon^2$$
$$^{\lambda/L\le 1}\frac{6\lambda}{25}\,|\,\widehat{w} \quad w_*\,|^2 + \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{32LF_*\log(2/\delta)}{\lambda n} + \left(\frac{32L^2}{\lambda} + \frac{416\lambda}{625}\right)\varepsilon^2.$$

By subtract ng λ $\widehat{w}$ w_* $^2/4$ fro both s des we co p ete the proof of _

Appendix E. Proof of Theorem 7

e cons der two cases_In the rst case we assu e that

$$\widehat{w} \quad w_* \quad \frac{1}{n^2}.$$

nce $H(\cdot)$ s L s ooth and $r(\cdot)$ s P L psch tz cont nuous we have

$$\begin{aligned} F(\widehat{w}) \quad F(w_*) &= H(\widehat{w}) + r(\widehat{w}) \quad H(w_*) \quad r(w_*) \\ &\quad \widehat{w} \quad w_*, \quad H(w_*) + \frac{L}{2} \widehat{w} \quad w_*^2 + P \widehat{w} \quad w_* \\ &\quad \widehat{w} \quad w_* \quad H(w_*) + \frac{L}{2} \widehat{w} \quad w_*^2 + P \widehat{w} \quad w_* \quad \frac{M+P}{n^2} + \frac{L}{2n^4} \end{aligned} \tag{4}$$

where the ast step ut zes Jensen s nequa ty

$$H(w_*) = \left\| E_{(x,y)\sim\mathbb{D}} [\quad (w_*, x, y)] \right\| \quad E_{(x,y)\sim\mathbb{D}} [\quad (w_*, x, y)] \quad M.$$

Next we study the case

$$\frac{1}{n^2} < \widehat{w} \quad w_* \quad 2R.$$

Fro 9 we have

$$\begin{aligned} &F(\widehat{w}) \quad F(w_*) + \frac{\lambda}{2} \widehat{w} \quad w_*^2 \\ &\quad F(\widehat{w}) \quad F(w_*) \quad [\widehat{F}(\widehat{w}) \quad \widehat{F}(w_*)], \widehat{w} \quad w_* + F(w_*) \quad \widehat{F}(w_*), \widehat{w} \quad w_* \\ &= H(\widehat{w}) \quad H(w_*) \quad [\widehat{H}(\widehat{w}) \quad \widehat{H}(w_*)], \widehat{w} \quad w_* + H(w_*) \quad \widehat{H}(w_*), \widehat{w} \quad w_* \\ &\quad \underbrace{\sup_{w:\|w-w_*\|\le\|\widehat{w}-w_*\|} \left\langle H(w) \quad H(w_*) \quad [\widehat{H}(w) \quad \widehat{H}(w_*)], w \quad w_* \right\rangle}_{:=B_1} \\ &\quad + \underbrace{\left\| H(w_*) \quad \widehat{H}(w_*) \right\|}_{:=B_2} \widehat{w} \quad w_* . \end{aligned}$$

(4)

e rst bound B_1_ o ut ze the fact the rando var ab e $\widehat{w}$ w_* es n the range $(1/n^2, 2R]$ we deve op the fo ow ng e a_

Lemma 5 *Under* **Assumptions 7** *and* **8**, *with probability at least* 1 δ, *for all*

$$\frac{1}{n^2} < y \quad 2R$$

the following bound holds:

$$\sup_{w:\|w-w_*\|\le\gamma} \left\langle H(w) \quad H(w_*) \quad [\widehat{H}(w) \quad \widehat{H}(w_*)], w \quad w_* \right\rangle \quad \frac{4Ly^2}{n} \left(8 + \sqrt{2 \log \frac{s}{\delta}} \right)$$

where $s = 2\log_2(n) + \log_2(2R)$.

Based on the above e a we have w th probab ty at east 1 δ

$$B_1 \quad \frac{4L\,|\,\widehat{w} \quad w_*\,|^2}{\bar{n}}\left(8+\sqrt{2\log\frac{s}{\delta}}\right)=\frac{LC\,|\,\widehat{w} \quad w_*\,|^2}{\bar{n}} \qquad 44$$

where C s de ned n _

e then proceed to hand e B_2 wh ch can be upper bounded n the sa e way as A_2_In part cu ar we have the fo ow ng e a_

Lemma 6 *Under* **Assumptions 7**, **8**, *and* **10**, *with probability at least* $1-\delta$, *we have*

$$\left\| H(w_*) \quad \widehat{H}(w_*)\right\| \quad \frac{2M\log(2/\delta)}{n}+\sqrt{\frac{8LH_*\log(2/\delta)}{n}}. \qquad 4$$

ubst tut ng 44 and 4 nto 4 w th probab ty at east 1 2δ we have

$$F(\widehat{w}) \quad F(w_*)+\frac{\lambda}{2}\,|\,\widehat{w} \quad w_*\,|^2$$
$$\frac{LC\,|\,\widehat{w} \quad w_*\,|^2}{\bar{n}}+\frac{2M\log(2/\delta)}{n}\,|\,\widehat{w} \quad w_*\,|+\,|\,\widehat{w} \quad w_*\,|\sqrt{\frac{8LH_*\log(2/\delta)}{n}}. \qquad 4$$

e subst tute

$$\frac{LC\,|\,\widehat{w} \quad w_*\,|^2}{\bar{n}} \quad \frac{L^2C^2\,|\,\widehat{w} \quad w_*\,|^2}{\lambda n}+\frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$\,|\,\widehat{w} \quad w_*\,|\sqrt{\frac{8LH_*\log(2/\delta)}{n}} \quad \frac{8LH_*\log(2/\delta)}{\lambda n}+\frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2$$

nto 4 and then have

$$F(\widehat{w}) \quad F(w_*) \quad \frac{L^2C^2\,|\,\widehat{w} \quad w_*\,|^2}{\lambda n}+\frac{2M\log(2/\delta)}{n}\,|\,\widehat{w} \quad w_*\,|+\frac{8LH_*\log(2/\delta)}{\lambda n}$$
$$\,|\,\frac{4R^2L^2C^2}{\lambda n}+\frac{4RM\log(2/\delta)}{n}+\frac{8LH_*\log(2/\delta)}{\lambda n}.$$

Co b n ng the above nequa ty w th 4 we obta n _

o prove , we subst tute

$$\frac{2M\log(2/\delta)}{n}\,|\,\widehat{w} \quad w_*\,| \quad \frac{8M^2\log^2(2/\delta)}{\lambda n^2}+\frac{\lambda}{8}\,|\,\widehat{w} \quad w_*\,|^2,$$
$$\,|\,\widehat{w} \quad w_*\,|\sqrt{\frac{8LH_*\log(2/\delta)}{n}} \quad \frac{16LH_*\log(2/\delta)}{\lambda n}+\frac{\lambda}{8}\,|\,\widehat{w} \quad w_*\,|^2$$

nto 4 and then have

$$F(\widehat{w}) \quad F(w_*)+\frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2$$
$$\frac{LC\,|\,\widehat{w} \quad w_*\,|^2}{\bar{n}}+\frac{8M^2\log^2(2/\delta)}{\lambda n^2}+\frac{16LH_*\log(2/\delta)}{\lambda n}$$
$$\frac{\lambda}{4}\,|\,\widehat{w} \quad w_*\,|^2+\frac{8M^2\log^2(2/\delta)}{\lambda n^2}+\frac{16LH_*\log(2/\delta)}{\lambda n}.$$

Co b n ng the above nequa ty w th 4 we obta n _

Appendix F. Proof of Lemma 5

F rst we part t on the range $(1/n^2, 2R]$ nto $s = \lceil 2\log_2(n) + \log_2(2R) \rceil$ consecut ve seg ents $\Delta_1, \Delta_2, \ldots, \Delta_s$ such that

$$\Delta_k = \left(\underbrace{\frac{2^{k-1}}{n^2}}_{:=\gamma_k^-}, \underbrace{\frac{2^k}{n^2}}_{:=\gamma_k^+} \right], \quad k = 1, \ldots, s.$$

hen we cons der the case $\gamma \quad \Delta_k$ for a xed va ue of k e have

$$\sup_{w:\|w-w_*\|\le\gamma} \Big\langle \quad H(w) \qquad H(w_*) \quad [\ \widehat{H}(w) \qquad \widehat{H}(w_*)], w \quad w_* \Big\rangle$$
$$\sup_{w:\|w-w_*\|\le\gamma_k^+} \Big\langle \quad H(w) \qquad H(w_*) \quad [\ \widehat{H}(w) \qquad \widehat{H}(w_*)], w \quad w_* \Big\rangle . \qquad 4$$

Based on the McD ar d s nequa ty McD ar d 9 9 and the ade acher co p ex ty Bart ett and Mende son 00 we have the fo ow ng e a to upper bound the ast ter _

Lemma 7 *Under* **Assumptions 7** *and* **8**, *with probability at least* $1 \quad \delta$, *we have*

$$\sup_{w:\|w-w_*\|\le\gamma_k^+} \Big\langle \quad H(w) \qquad H(w_*) \quad [\ \widehat{H}(w) \qquad \widehat{H}(w_*)], w \quad w_* \Big\rangle$$
$$\frac{L(\gamma_k^+)^2}{n}\left(8 + \sqrt{2\log\frac{1}{\delta}}\right). \qquad 4$$

nce $\gamma \quad \Delta_k$ we have

$$\gamma_k^+ = 2\gamma_k^- \quad 2\gamma. \qquad 49$$

hus w th probab ty at east $1 \quad \delta$ we have

$$\sup_{w:\|w-w_*\|\le\gamma} \Big\langle \quad H(w) \qquad H(w_*) \quad [\ \widehat{H}(w) \qquad \widehat{H}(w_*)], w \quad w_* \Big\rangle$$
$$4 \quad 4 \quad 49 \quad \frac{4L\gamma^2}{n}\left(8 + \sqrt{2\log\frac{1}{\delta}}\right).$$

e co p ete the proof by ta ng the un on bound over s seg ents_

Appendix G. Proof of Lemma 7

o s p fy the notat on we de ne

$$h_i(w) = \ell(w, x_i, y_i), \quad i = 1, \ldots, n,$$
$$\Gamma(h_1, \ldots, h_n) = \sup_{w:\|w-w_*\|\le\gamma_k^+} \Big\langle \quad H(w) \qquad H(w_*) \quad \frac{1}{n}\sum_{i=1}^{n}[\ h_i(w) \qquad h_i(w_*)], w \quad w_* \Big\rangle .$$

o upper bound $\Gamma(h_1, \ldots, h_n)$ we ut ze the McD ar d s nequa ty McD ar d 9 9 _

Theorem 8 *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $H : A^n$ $\mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} \left| H(x_1, \ldots, x_n) \quad H(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \quad c_i$$

for every $1 \quad i \quad n$. *Then, for every* $t > 0$,

$$P \cdot H(X_1, \ldots, X_n) \quad E[H(X_1, \ldots, X_n)] \quad t \quad \exp\left(\frac{2t^2}{\sum_{i=1}^n c_i^2} \right).$$

As po nted out n **Remark 7** **Assumptions 7** and **8** p y the rando funct on $h_i(\cdot)$ s L s ooth and thus

$$\quad h_i(w) \quad h_i(w_*), w \quad w_* \quad L \, w \quad w_* \, ^2 \quad L \left(\gamma_k^+\right)^2.$$

As a resu t when a rando funct on h_i changes the rando var ab e $I(h_1, \ldots, h_n)$ can change by no ore than $2L\left(\gamma_k^+\right)^2 / n$ o see th s we have

$$I(h_1, \ldots, h_n) \quad I(h_1, \ldots, h_{i-1}, h_i', h_{i+1}, \ldots, h_n)$$
$$\frac{1}{n} \sup_{w : \|w - w_*\| \le \gamma_k^+} \langle \quad h_i'(w) \quad h_i'(w_*) \quad [\quad h_i(w) \quad h_i(w_*)], w \quad w_* \rangle \quad \frac{2}{n} L \left(\gamma_k^+\right)^2.$$

McD ar d s nequa ty p es that w th probab ty at east $1 \quad \delta$

$$I(h_1, \ldots, h_n) \quad E[I(h_1, \ldots, h_n)] + L \left(\gamma_k^+\right)^2 \sqrt{\frac{2}{n} \log \frac{1}{\delta}}. \qquad \blacksquare$$

Let $(h_1', \ldots, h_n')$ be an ndependent copy of $(h_1, \ldots, h_n)$ and $_1, \ldots, _n$ be n d ade a cher var ab es w th equa probab ty of be ng 1 s ng techn ques of ade acher co p ex t es Bart ett and Mende son we bound $E[I(h_1, \ldots, h_n)]$ as fo ows·

$$E_{h_1,\ldots,h_n} \left[\sup_{w : \|w - w_*\| \le \gamma_k^+} \left\langle \quad H(w) \quad H(w_*) \quad \frac{1}{n} \sum_{i=1}^n [\quad h_i(w) \quad h_i(w_*)], w \quad w_* \right\rangle \right]$$
$$= \frac{1}{n} E_{h_1,\ldots,h_n} \left[\sup_{w : \|w - w_*\| \le \gamma_k^+} \right.$$
$$\left. E_{h_1',\ldots,h_n'} \left[\sum_{i=1}^n \langle \quad h_i'(w) \quad h_i'(w_*), w \quad w_* \rangle \right] \quad \sum_{i=1}^n \quad h_i(w) \quad h_i(w_*), w \quad w_* \right]$$
$$\frac{1}{n} E_{h_1,\ldots,h_n,h_1',\ldots,h_n'} \left[\sup_{w : \|w - w_*\| \le \gamma_k^+} \right.$$
$$\left. \sum_{i=1}^n \langle \quad h_i'(w) \quad h_i'(w_*), w \quad w_* \rangle \quad \sum_{i=1}^n \quad h_i(w) \quad h_i(w_*), w \quad w_* \right]$$

$$= \frac{1}{\mathbf{n}} \mathrm{E}_{h_1,\ldots,h_n,h'_1,\ldots,h'_n,\epsilon_1,\ldots,\epsilon_n} \left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+} \right.$$

$$\sum_i^n$$

Note that $\mathbf{x}^2$ s 2a L psch tz over [a, a] and $\mathbf{p}_i(\mathbf{w})+\mathbf{q}_i(\mathbf{w})$ [$2\gamma_k^+\mathbf{D}$ $\bar{\beta}$, $2\gamma_k^+\mathbf{D}$ $\bar{\beta}$]_ hen fro the co par son theore of ade acher co p ex t es Ledoux and a agrand 99 n part cu ar Le a of Me r and Zhang 00 we have

$$\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\left(\mathbf{p}_i(\mathbf{w})+\mathbf{q}_i(\mathbf{w})\right)^2\right]$$
$$4\gamma_k^+\mathbf{D}\sqrt{\beta}\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\left(\mathbf{p}_i(\mathbf{w})+\mathbf{q}_i(\mathbf{w})\right)\right]$$
$$4\gamma_k^+\mathbf{D}\sqrt{\beta}\left(\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\mathbf{p}_i(\mathbf{w})\right]+\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\mathbf{q}_i(\mathbf{w})\right]\right).$$

ar y we have

$$\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\left(\mathbf{p}_i(\mathbf{w})\quad\mathbf{q}_i(\mathbf{w})\right)^2\right]$$
$$4\gamma_k^+\mathbf{D}\sqrt{\beta}\left(\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\mathbf{p}_i(\mathbf{w})\right]+\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\mathbf{q}_i(\mathbf{w})\right]\right). \quad 4$$

Co b n ng and 4 we arr ve at

$$\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\ \ \mathbf{h}_i(\mathbf{w})\quad\mathbf{h}_i(\mathbf{w}_*),\mathbf{w}\quad\mathbf{w}_*\right]$$
$$2\gamma_k^+\mathbf{D}\sqrt{\beta}\left(\underbrace{\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\mathbf{p}_i(\mathbf{w})\right]}_{:=C_1}+\underbrace{\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\mathbf{q}_i(\mathbf{x})\right]}_{:=C_2}\right).$$

e proceed to upper bound $\mathbf{C}_1$ n _Fro our de n t on of $\mathbf{p}_i(\mathbf{w})$ we have

$$\left|\mathbf{p}_i(\mathbf{w})\quad\mathbf{p}_i(\mathbf{w}')\right|=\frac{1}{\beta}\left|\_'(\ \mathbf{w},\mathbf{x}_i\ ,\mathbf{y}_i)\quad\_'(\ \mathbf{w}',\mathbf{x}_i\ ,\mathbf{y}_i)\right|$$
$$\sqrt{\beta}\left|\ \mathbf{w},\mathbf{x}_i\quad\ \mathbf{w}',\mathbf{x}_i\ \right|=\sqrt{\beta}\left|\ \mathbf{x}_i,\mathbf{w}\quad\mathbf{w}_*\quad\ \mathbf{x}_i,\mathbf{w}'\quad\mathbf{w}_*\right|.$$

App y ng the co par son theore of ade acher co p ex t es aga n we have

$$\mathbf{C}_1\quad\sqrt{\beta}\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n \_i\ \mathbf{x}_i,\mathbf{w}\quad\mathbf{w}_*\right]=\mathbf{C}_2.$$