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 $\tilde{}(\quad + \sqrt{\quad *})$ r s bound when athrO

where $= \cdot : \mathcal{X} \quad \mathbb{R}\bullet$ s a hypothes s c ass $(\mathbf{x} \quad) \quad \mathcal{X} \quad \mathbb{R}$ s an nstance abe pa r sa p ed fro a d str but on $\mathbb{D}$ and $(\_\_) : \mathbb{R} \quad \mathbb{R} \quad \mathbb{R}$ 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 $(\_)$ 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 8 h e both A and E M have been extens ve y stud ed n recent years Bart ett and Mende 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 funct ons $_1 \quad _n$ sa p ed ndependent y fro $\mathbb{P}$ and a to n ze an e p r ca ob ect ve funct on

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

Let $\widehat{\mathbf{w}}$ $\operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{\ }(\mathbf{w})$ 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

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

tate of the art r s bounds of E M nc ude an $\widetilde{\ }(\sqrt{\quad})$ bound when the rando funct on $(\_)$ s L psch tz cont nuous where s the d ens ona ty of $\mathbf{w}$, an $(1 \quad)$ bound when $(\_)$ s strong y convex ha ev hwartz et a 9 , and an $\widetilde{\ }(\quad)$ bound when $(\_)$ s exponent a y concave 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 $_* = \min_{\mathbf{w}\in\mathcal{W}} \quad (\mathbf{w})$ be the n a r s be the odu us of strong convex ty of $(\_)$ and be the odu us of s oothness of $(\_)$ Denote by $=$ 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 $(\_)$ the e p r ca funct on $\widehat{\ }(\mathbf{w})$ and the expected funct on $(\_)$ For our resu ts of E M for CO we assu e the do a n s bounded and the rando funct on s nonegat 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 exponent a y concave exp

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

hen ( ) s both convex and s ooth and ( ) s L psch tz cont nuous we estab sh an $\widetilde{}($ $+ \sqrt{\ {}_*}$ $)$ r s bound c f heore In the opt st c case that $_*$ s s a e $_* =$ (^{2}), we obta n an $\widetilde{}($ $)$ r s bound wh ch s ana ogous to the $\widetilde{}(1$ $)$ opt st c rate of E M for superv sed earn ng rebro et a If ( ) s a so strong y convex we prove an $\widetilde{}($ $+$ $_*$ $)$ r s bound and prove t to (1 [^{2}] + $_*$) when $= \widetilde{\Omega}($ $)$ c f heore hus f s arge and $_*$ s s a e $_* =$ (1) we get an (^{2}) r s bound wh ch to the best of our now edge s the rst (1 ^{2}) type of r s bound of E M

hen convex ty s not present n ( ) as ong as ( ) s s ooth $\widehat{}$( ) s convex and ( ) s strong y convex we st obta n an proved r s bound of (1 [^{2}] + $_*$) when $=$ $\widetilde{\Omega}($ ^{2} $)$ wh ch w further p y an (^{2}) r s bound f $_* =$ (1) c f heore

F na y we extend the (1 [^{2}] + $_*$) 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 can be rep aced w th $\Omega($ $^2)$ 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 ( ) s L psch tz cont nuous ha ev hwartz et a 9 de onstrate that $\widehat{\ }(\mathbf{w})$ converges to $(\mathbf{w})$ un for y over w th an $\widetilde{\ }(\sqrt{\ })$ error bound that ho ds w th h gh probab ty p y ng an $\widetilde{\ }(\sqrt{\ })$ r s bound of E M hey further estab sh an (1) r s bound of E M that ho ds n expectat on when ( ) s 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 and Mehta proves an $\widetilde{\ }($ $)$ bound of E M that ho ds w th h gh probab ty when ( ) s 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 who exh b ts a ower bound of $\Omega(\ ^2)$ sa p e co p ex ty for un for convergence that near y atches the upper bound of ha ev hwartz et a 9 , and a ower bound of $\Omega(\ )$ sa p e co p ex ty of E M wh ch s atched by our $\widetilde{\ }(\ + \sqrt{\ _*})$ bound when $_*$ 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 $\widehat{\ }(\ )$ to $(\ )$ over the hypothes s c ass Ko tch ns In fact un for convergence s a suf c ent cond t on for earnab ty ha ev hwartz and Ben Dav d 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 8 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 $(\sqrt{\mathrm{VC}(\ )\ })$ where $\mathrm{VC}(\ )$ s the C d ens on of If the oss ( ) s L psch tz con t nuous w th respect to ts rst argu ent we have a r s bound of $(1\ \ + \ _n(\ ))$ where $_n(\ )$ s the ade acher co p ex ty of he ade acher co p ex ty typ ca y sca es as $_n(\ ) = (1\ \ )$ e g conta ns near funct ons w th ow nor p y ng an $(1\ \ )$ r s bound Bart ett and Mende son here have been ntens ve efforts to der ve rates faster than $(1\ \ )$ under var ous cond t ons Lee et a 99 , Panchen o , Bart ett et a , Gonen and ha ev hwartz such as ow no se syba ov 4 s oothness rebro et a strong convex ty r dharan et a 9 to na e a few a ongst any pec ca y when the rando funct on ( ) s nonnegat ve and s ooth rebro et a have estab shed a r s bound of $\widetilde{\ }(\ ^2_n(\ ) + \ _n(\ )\ \ _*)$ reduc ng to an $\widetilde{\ }(1\ )$ bound f $_n(\ ) = (1\ \ )$ and $_* = (1\ )$ A genera zed near for of s stud ed by r dharan et a 9 and a r s bound of (1) s proved f the expected funct on ( ) s 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 to denote the $_2$ nor of vectors

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

$$\mathbf{w} \leq \quad \mathbf{w}$$

Assumption 2 *The random function* $(\cdot)$ *is nonnegative, and -smooth over , that is,*

$$\| \quad (\mathbf{w}) \qquad (\mathbf{w}')\| \leq \quad \mathbf{w} \quad \mathbf{w}' \quad \mathbf{w} \ \mathbf{w}' \qquad \mathbb{P}$$

Assumption 3 *The expected function* $(\cdot)$ *is -Lipschitz continuous over , that is,*

$$(\mathbf{w}) \qquad (\mathbf{w}') \leq \quad \mathbf{w} \quad \mathbf{w}' \quad \mathbf{w} \ \mathbf{w}'$$

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

(a) *The expected function* $(\cdot)$ *is convex over .*

(b) *The expected function* $(\cdot)$ *is -strongly convex over , that is,*

$$(\mathbf{w}) + \quad (\mathbf{w}) \ \mathbf{w}' \quad \mathbf{w} + \frac{}{2} \mathbf{w}' \quad \mathbf{w}^2 \leq \quad (\mathbf{w}') \quad \mathbf{w} \ \mathbf{w}'$$

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

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

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

$$(\mathbf{w}_*) \leq \qquad \mathbb{P}$$ 8

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 $(\cdot)$ p es the expected funct on $(\cdot)$ s s ooth By Jensen s nequa ty we have

$$\| \quad (\mathbf{w}) \qquad (\mathbf{w}')\| \leq \mathrm{E}_{f \sim \mathbb{P}} \| \quad (\mathbf{w}) \qquad (\mathbf{w}')\| \leq \quad \mathbf{w} \quad \mathbf{w}' \quad \mathbf{w} \ \mathbf{w}'$$

ar y the e p r ca funct on $\widehat{\ }(\cdot)$ s a so s ooth he *condition number* of $(\cdot)$ s de ned as the rat o between and e $=$ 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 1 2, 0, *define*

$$(\quad) = 2\left(\log \frac{2}{} + \quad \log \frac{6}{}\right) \qquad (9)$$

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

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*)$$
$$< \frac{16 \quad ^2 \quad (\quad)}{} + \frac{8 \quad \log(2 \quad)}{} + 8 \sqrt{\frac{2 \quad _* \log(2 \quad)}{}} + \left(8 \quad + \quad + \frac{4 \quad (\quad)}{}\right)$$

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

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

Corollary 2 *By setting* $= 1$ *in Theorem 1, we have* $(1 \quad) = 2\left(\log(2 \quad) + \log(6 \quad)\right)$, *and with high probability*

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) = \left(\frac{\log}{} + \sqrt{\frac{*}{}}\right) = \widetilde{\ } \left(- + \sqrt{\frac{*}{}}\right)$$

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 $\widetilde{\ }(\quad + \sqrt{\ _*\ })$ r s bound for CO hen the n a r s s s a e $_* = (\ ^2 \ )$ the rate s proved to $\widetilde{\ }(\quad)$ Note that even under the s oothness assu pt on the near dependence on 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* $1 \quad 2$, *we have*

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*)$$
$$< \frac{16 \quad ^2 \quad (\quad)}{} + \frac{8 \quad \log(2 \quad)}{} + \frac{8 \quad _* \log(2 \quad)}{} + \left(8 \quad + \quad + \frac{4 \quad (\quad)}{}\right)$$

Furthermore, if

$$\frac{4 \quad (\quad)}{} = 4 \quad (\quad)$$

we also have

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) < \frac{32 \quad ^2 \log^2(2 \quad)}{^2} + \frac{128 \quad _* \log(2 \quad)}{} + \left(\frac{128 \quad ^2 \ ^2}{} + 16 \quad + 4 \quad ^2\right)$$

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

Corollary 4 *By setting* $= 1$ *in Theorem 3, we have* $(1 \quad) = 2\left(\log(2 \quad) + \log(6 \quad)\right)$, *and with high probability*

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) = \left(\frac{\log}{} + \frac{*}{}\right) = \widetilde{\ } \left(- + \frac{*}{}\right)$$

By setting $= 1 \quad ^2$, *we have* $(1 \quad ^2 \quad) = 2\left(\log(2 \quad) + \log(6 \quad ^2 \quad)\right)$ *and when* $= \Omega(\quad \log \quad) = \widetilde{\Omega}(\quad)$, *with high probability*

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) = \left(\frac{1}{^2} + \frac{*}{}\right)$$

Remark 3 he rst part of Coro ary **4** shows that E M en oys an $\tilde{}$ (+ $_*$) 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 (1) r s bound proved by ha ev hwartz et a 9 but w th str ng d fferences h gh ghted n ab e nce the r s bound of ha ev hwartz et a 9 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 s ncreased by a factor of 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 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 $(1 [\;^2] + \; _*)$ r s bound for superv sed earn ng However a a or tat on of these theore s s that the ower bound of depends on the d ens ona ty and thus cannot be app ed to n n te d ens ona cases e g erne ethods chò 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

$$\min_{\mathbf{w}\in\mathcal{W}} \quad (\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}} [\;(\mathbf{w} \;\; \mathbf{x} \quad)] + \; (\mathbf{w})$$

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

$$\min_{\mathbf{w}\in\mathcal{W}} \;\widehat{\;}(\mathbf{w}) = \frac{1}{\;}\sum_{i=1}^{n} \;(\mathbf{w} \;\; \mathbf{x}_i \quad _i) + \; (\mathbf{w})$$

e de ne

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

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

$$\left\| (\mathbf{w}_* \; \mathbf{x} \quad) \right\| < \quad (\mathbf{x} \quad) \quad \mathbb{D}$$

Remark 7 he above assu pt ons a ow us to ode any popu ar osses n ach ne earn ng such as regu ar zed square oss and regu ar zed og st c oss **Assumptions 7** and **8** p y the rando funct on (_x) s ^{2} s ooth over o see th s for any w w$'$ we have

$$\begin{aligned} \| (\mathbf{w} \; \mathbf{x} \quad) \quad (\mathbf{w}' \; \mathbf{x} \quad)\| &= \| \,'(\mathbf{w} \; \mathbf{x} \quad)\mathbf{x} \quad '(\mathbf{w}' \; \mathbf{x} \quad)\mathbf{x}\| \\ \overset{9}{<} \quad '(\mathbf{w} \; \mathbf{x} \quad) \quad '(\mathbf{w}' \; \mathbf{x} \quad) &< \quad \mathbf{w} \; \mathbf{x} \quad \mathbf{w}' \; \mathbf{x} \overset{9}{<} \quad {}^2 \|\mathbf{w} \quad \mathbf{w}'\| \end{aligned}$$

By Jensen s nequa ty ( ) s a so ^{2} s ooth Not ce that ^{2} s the odu us of s oothness of ( ) and s the odu us of strong convex ty of ( ) th a s ght abuse of notat on we de ne $=$ ^{2} and the cond t on nu ber as the rat o between and e $=$ F na y we note that the regu ar zer ( ) cou d be *non-smooth*

e have the fo ow ng excess r s bound of E M for superv sed earn ng

Theorem 7 *For any* 0 1 2*, define*

$$= 4\left(8 + \sqrt{2\log \frac{2\log_2(\;) + \log_2(2 \;)}{\quad}}\right)$$

$$_* = \quad (\mathbf{w}_*) = \quad (\mathbf{w}_*) \quad (\mathbf{w}_*)$$

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

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) < \max\left(\frac{+}{2} + \frac{}{2 \quad 4} \frac{4 \quad {}^2 \quad {}^2 \quad {}^2}{} + \frac{4 \quad \log(2 \;)}{} + \frac{8 \quad {}_* \log(2 \;)}{}\right)$$

Furthermore, if

$$\frac{16 \quad {}^2 \quad {}^2}{2} = 16 \quad {}^2 \quad {}^2$$

with probability at least 1 2 *, we have*

$$(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) < \max\left(\frac{+}{2} + \frac{}{2 \quad 4} \frac{8 \quad {}^2\log^2(2 \;)}{2} + \frac{16 \quad {}_* \log(2 \;)}{}\right)$$

Remark 8 he rst part of heore presents an ( ) r s bound s ar to the (1) r s bound of r dharan et a 9 he second part s an (1 [^{2}] + $_*$) r s bound and n th s case the ower bound of s $\Omega($ $^2)$ wh ch s d ens ona ty ndependent hus heore can be app ed even when the d ens ona ty s n n te Genera y spea ng the regu ar zer ( ) s nonnegat ve and thus $_* <$ $_*$ o the second bound s even better than those n heore s and

F na y we note that heore shou d be treated as a counterpart of heore for superv sed earn ng because both of the do not re y on the nd v dua convex ty e **Assumption 4(d)** One ay wonder whether t s poss b e to der ve a counterpart of heore that s whether t s poss b e to ut ze the nd v dua convex ty to reduce the ower bound of by a factor of e w nvest gate th s quest on as a future wor

For brev ty we treat C as a constant because t on y has a *double* ogar th c 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 $\widehat{\ }(\_)$ and the opt a ty cond t on of $\widehat{\mathbf{w}}$ Boyd and andenberghe 4 we have

Lemma 1 *Under* **Assumptions 2** *and* **4(d)***, with probability at least* 1 , *w024(fo83399(b)0.0r.841(w4)0.510254(n)0* where the ast step s due to

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \sqrt{\frac{(\ )(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*))}{}} < \frac{(\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{2} + \frac{(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*)}{2}$$

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \sqrt{\frac{(\ )}{}} < \frac{(\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{2} + \frac{}{2}$$

Fro 4 we get

$$\begin{aligned} &\frac{1}{2}(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*)) \\ < &\frac{2 \quad (\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \frac{2 \quad \log(2 \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \sqrt{\frac{8 \quad _* \log(2 \ )}{}} \\ &+ 2 \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| + \frac{}{2} + \frac{(\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} \\ \overset{4}{<} &\frac{8 \ ^2 \quad (\ )}{} + \frac{4 \quad \log(2 \ )}{} + 4 \sqrt{\frac{2 \quad _* \log(2 \ )}{}} + \left(4 \quad + \frac{}{2} + \frac{2 \quad (\ )}{}\right) \end{aligned}$$

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 (1) 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 s arge enough E M has (1 ^{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 $_2$ 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 $\widetilde{\ }$( ) or $\widetilde{\ }$(1) 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 n the second part of heore Idea y we expect that $= \Omega(\ )$ s suf c ent to de ver an (1 [^{2}] + $_*$) r s bound

4 he (1 ^{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 (1 ^{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 Jangsu F BK 8 N F II 4 9 II 4 99 and the Co aborat ve Innovat on 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 convergence rate (1) 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* arX 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* arX v 4

E ad Hazan and atyen Ka e Beyond the regret n zat on barr er an opt a a gor th for stochast 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 n *Stochastic Approximation and Recursive Algorithms and Applications* pr nger J 9 4 d e 9 9 9 4 ne 49 9 te 4 w o

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 4 8 88 8 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* 4 8 9 8

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 4 4 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 4 4 9 9

Yur Nesterov *Introductory lectures on convex optimization: a basic course* vo u e 8 of *Applied optimization* K uwer Acade c Pub shers 4

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 8 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* 8 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 449 4

Bernhard chö 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 4

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 uszczyns *Lectures on Stochastic Programming: Modeling and Theory* IAM second ed t on 4

teve a e and D ng Xuan Zhou Learn ng theory est ates v a ntegra operators and the r approx 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 8

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 8 9 88 a

L un Zhang anbao Yang ong J n and X aofe He (log) 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 be a random variable with values in . Assume $\| \| <$ almost surely. Denote $^2( ) = \mathrm{E}[\| \|^2]$. Let $\{ _i\}_{i=1}^m$ be ( ) independent drawers of . For any $0 < < 1$, with confidence 1 ,*

$$\left\| \frac{1}{m}\sum_{i=1}^{m}[_i - \mathrm{E}[_i]] \right\| < \frac{2 \log(2/)}{ } + \sqrt{\frac{2 ^2( )\log(2/)}{ }}$$

e rst cons der a xed w ( ) nce $_i(\cdot)$ s s ooth we have

$$\| _i(\mathbf{w}) - _i(\mathbf{w}_*)\| < \|\mathbf{w} - \mathbf{w}_*\|$$

Because $_i(\cdot)$ s both convex and s ooth by 4 of Nesterov 4 we have

$$\| _i(\mathbf{w}) - _i(\mathbf{w}_*)\|^2 < (_i(\mathbf{w}) - _i(\mathbf{w}_*) - \langle _i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_*\rangle)$$

a ng expectat on over both s des we have

$$\mathrm{E}\left[\| _i(\mathbf{w}) - _i(\mathbf{w}_*)\|^2\right] < ((\mathbf{w}) - (\mathbf{w}_*) - \langle (\mathbf{w}_*), \mathbf{w} - \mathbf{w}_*\rangle) < ((\mathbf{w}) - (\mathbf{w}_*))$$

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

$$\langle (\mathbf{w}_*), \mathbf{w} - \mathbf{w}_*\rangle \geq 0, \forall \mathbf{w}$$

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

$$\begin{aligned}
&\left\| (\mathbf{w}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\mathbf{w}) \quad \widehat{\ }(\mathbf{w}_*)]\right\| \\
&= \left\| (\mathbf{w}) \quad (\mathbf{w}_*) \quad \frac{1}{\ }\sum_{i=1}^{n}[\ _i(\mathbf{w}) \quad _i(\mathbf{w}_*)]\right\| \\
&\leq \frac{2 \,\|\mathbf{w} \quad \mathbf{w}_*\| \log(2\ \ )}{\ } + \sqrt{\frac{2\ (\ (\mathbf{w}) \quad (\mathbf{w}_*))\log(2\ \ )}{\ }}
\end{aligned}$$

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

Let $\mathcal{B}$ be an un t ba of d ens on and $(\mathcal{B}\ \ )$ be ts 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

$$\log (\mathcal{B}\ \ ) \leq \log \frac{3}{\ }$$

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

$$\log (\ \ ) \leq \log \left| \left(\mathcal{B}(\ )\ \frac{\ }{2}\right)\right| \leq \log \frac{6}{\ }$$

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 $\mathrm{E}\left[\|\ _i(\mathbf{w}_*)\|^2\right]$ nce $_i(\cdot)$ s s ooth and nonnegat ve fro Le a 4 of rebro et a we have

$$\|\ _i(\mathbf{w}_*)\|^2 \leq 4\ \ _i(\mathbf{w}_*)$$

and thus

$$\mathrm{E}\left[\|\ _i(\mathbf{w}_*)\|^2\right] \leq 4\ \mathrm{E}[\ _i(\mathbf{w}_*)] = 4\ \ _*$$

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

$$\left\| (\mathbf{w}_*) \quad \widehat{\ }(\mathbf{w}_*)\right\| = \left\| (\mathbf{w}_*) \quad \frac{1}{\ }\sum_{i=1}^{n}\ _i(\mathbf{w}_*)\right\| \leq \frac{2 \quad \log(2\ \ )}{\ } + \sqrt{\frac{8 \quad _* \log(2\ \ )}{\ }}$$

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

$$\begin{aligned}
&(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) + \frac{\ }{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
&\leq \Bigg(\underbrace{\left\| (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\widehat{\mathbf{w}}) \quad \widehat{\ }(\mathbf{w}_*)]\right\|}_{:=A_1} + \underbrace{\left\| (\mathbf{w}_*) \quad \widehat{\ }(\mathbf{w}_*)\right\|}_{:=A_2}\Bigg)\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|
\end{aligned} \quad (8)$$

ubst tut ng and nto 8 w th probab ty at east 1 2 we have

$$
\begin{aligned}
&(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) + \frac{}{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
&< \frac{(\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ )(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*))}{}} \\
&+ \frac{2 \quad \log(2\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{8 \quad {}_* \log(2\ \ )}{}} \\
&+ 2 \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| + \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ )}{}} + \frac{(\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{}
\end{aligned}
\qquad 9
$$

o prove we subst tute and

$$
\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{8 \quad {}_* \log(2\ \ )}{}} < \frac{4 \quad {}_* \log(2\ \ )}{} + \frac{}{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2
$$

nto 9 and then obta n

$$
\begin{aligned}
&\frac{1}{2}(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*)) \\
&< \frac{2 \quad (\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \frac{2 \quad \log(2\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \frac{4 \quad {}_* \log(2\ \ )}{} \\
&+ 2 \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| + \frac{}{2} + \frac{(\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} \\
&< \frac{8\ ^2 \quad (\ \ )}{} + \frac{4 \quad \log(2\ \ )}{} + \frac{4 \quad {}_* \log(2\ \ )}{} + \left(4 \quad + \frac{}{2} + \frac{2 \quad (\ \ )}{}\right)
\end{aligned}
$$

wh ch p es

o prove we subst tute

$$
\begin{aligned}
\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ )(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*))}{}} &< \frac{2 \quad (\ \ )(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*))}{} + \frac{}{8}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
\frac{2 \quad \log(2\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} &< \frac{16 \quad ^2 \log^2(2\ \ )}{^2} + \frac{}{16}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{8 \quad {}_* \log(2\ \ )}{}} &< \frac{64 \quad {}_* \log(2\ \ )}{} + \frac{}{32}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
2 \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| &< \frac{64 \quad ^2\ ^2}{} + \frac{}{64}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ )}{}} &< \frac{32 \quad (\ \ )}{} + \frac{}{128}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
\frac{(\ \ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} &< \frac{32 \quad ^2 \quad ^2(\ )\ ^2}{^2} + \frac{}{128}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2
\end{aligned}
$$

nto (9) and then obta n

$$
\begin{aligned}
&(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) + \frac{}{4}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
&< \frac{(\ \ ,)\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \frac{2 \quad (\ \ ,)(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*))}{} + \frac{16 \quad ^2\log^2(2\ \ ,)}{^2} + \frac{64 \quad _*\log(2\ \ ,)}{} \\
&\quad + \frac{64 \ ^2\ ^2}{} + \frac{32 \quad (\ \ ,)}{} + \frac{32 \ ^2\ \ ^2(\ ) \ ^2}{^2} \\
&< \frac{}{4}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 + \frac{1}{2}(\ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*)) + \frac{16 \quad ^2\log^2(2\ \ ,)}{^2} + \frac{64 \quad _*\log(2\ \ ,)}{} \\
&\quad + \frac{64 \ ^2\ ^2}{} + 8 \quad + 2 \quad ^2
\end{aligned}
$$

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 , *for any* $\mathbf{w}$ ( ), *we have*

$$
\left\| \quad (\mathbf{w}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\mathbf{w}) \quad \widehat{\ }(\mathbf{w}_*)]\right\| < \frac{(\ \ ,)\|\mathbf{w} \quad \mathbf{w}_*\|}{} + \ \|\mathbf{w} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ ,)}{}}
$$

where (,) *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 we have

$$
\begin{aligned}
&\left\| \quad (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\widehat{\mathbf{w}}) \quad \widehat{\ }(\mathbf{w}_*)]\right\| \\
&< \frac{(\ \ ,)\|\widetilde{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \ \|\widetilde{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ ,)}{}} + 2 \\
&< \frac{(\ \ ,)\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ ,)}{}} + \frac{(\ \ ,)}{} + \quad \sqrt{\frac{(\ \ ,)}{}} + 2
\end{aligned} \quad (4)
$$

ubst tut ng (4) and nto (8) w th probab ty at east 1 2 we have

$$
\begin{aligned}
&(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) + \frac{}{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
&< \frac{(\ \ ,)\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2\sqrt{\frac{(\ \ ,)}{}} \\
&\quad + \frac{2 \quad \log(2\ \ ,)\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{8 \quad _*\log(2\ \ ,)}{}} \\
&\quad + 2 \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| + \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ \ ,)}{}} + \frac{(\ \ ,)\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{}
\end{aligned} \quad (4)
$$

o get 4 we subst tute

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2\sqrt{\frac{(\ ,\ )}{}} \lesssim \frac{^2\ (\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \frac{}{4}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{8 \quad _* \log(2\ ,\ )}{}} \lesssim \frac{8 \quad _* \log(2\ ,\ )}{} + \frac{}{4}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

nto 4 and then obta n

$$\begin{aligned}
&(\widehat{\mathbf{w}}) \qquad (\mathbf{w}_*) \\
&\lesssim \frac{(\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \frac{^2\ (\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \frac{2 \quad \log(2\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} + \frac{8 \quad _* \log(2\ ,\ )}{} \\
&\quad + 2 \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| + \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ ,\ )}{}} + \frac{(\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} \\
&\overset{4}{\lesssim} \frac{4 \quad ^2 \quad (\ ,\ )}{} + \frac{4 \quad ^2 \quad ^2 \quad (\ ,\ )}{} + \frac{4 \quad \log(2\ ,\ )}{} + \frac{8 \quad _* \log(2\ ,\ )}{} \\
&\quad + \left(4 \quad + 2 \quad \sqrt{\frac{(\ ,\ )}{}} + \frac{2 \quad (\ ,\ )}{}\right)
\end{aligned}$$

wh ch proves 4

o get we subst tute

$$\frac{2 \quad \log(2\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} \lesssim \frac{8 \quad ^2 \log^2(2\ ,\ )}{^2} + \frac{}{8}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{8 \quad _* \log(2\ ,\ )}{}} \lesssim \frac{32 \quad _* \log(2\ ,\ )}{} + \frac{}{16}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

$$2 \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \lesssim \frac{32 \quad ^2 \quad ^2}{} + \frac{}{32}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|\sqrt{\frac{(\ ,\ )}{}} \lesssim \frac{16 \quad ^2 \quad (\ ,\ )^2}{} + \frac{}{64}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

$$\frac{(\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|}{} \lesssim \frac{16 \quad ^2 \quad ^2(\ )^2}{^2} + \frac{}{64}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$$

nto 4 and then obta n

$$\begin{aligned}
&(\widehat{\mathbf{w}}) \qquad (\mathbf{w}_*) + \frac{}{4}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
&\lesssim \frac{(\ ,\ )\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{} + \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2\sqrt{\frac{(\ ,\ )}{}} + \frac{8 \quad ^2 \log^2(2\ ,\ )}{^2} + \frac{32 \quad _* \log(2\ ,\ )}{} \\
&\quad + \left(\frac{32 \quad ^2}{} + \frac{16 \quad ^2 \quad (\ ,\ )}{} + \frac{16 \quad ^2 \quad ^2(\ )}{^2}\right) \quad ^2 \\
&\lesssim \frac{^2\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2}{25} + \frac{}{5}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 + \frac{8 \quad ^2 \log^2(2\ ,\ )}{^2} + \frac{32 \quad _* \log(2\ ,\ )}{} \\
&\quad + \left(\frac{32 \quad ^2}{} + \frac{16}{25} + \frac{16 \quad ^3}{625 \quad ^2}\right) \quad ^2 \\
&\overset{\lambda/L\leq 1}{\lesssim} \frac{6}{25}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 + \frac{8 \quad ^2 \log^2(2\ ,\ )}{^2} + \frac{32 \quad _* \log(2\ ,\ )}{} + \left(\frac{32 \quad ^2}{} + \frac{416}{625}\right) \quad ^2
\end{aligned}$$

By subtract ng $\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2$ 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

$$\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| < \frac{1}{\ \ 2}$$

nce ( ) s s ooth and ( ) s L psch tz cont nuous we have

$$
\begin{aligned}
(\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) &= \quad (\widehat{\mathbf{w}}) + \ (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) \quad (\mathbf{w}_*) \\
&\leq \widehat{\mathbf{w}} \quad \mathbf{w}_* \quad (\mathbf{w}_*) \ + \frac{}{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 + \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \\
&\leq \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \| \quad (\mathbf{w}_*)\| + \frac{}{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 + \ \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| < \frac{\ \ +\ \ }{\ \ 2} + \frac{}{2 \ \ 4}
\end{aligned}
$$

where the ast step ut zes Jensen s nequa ty

$$\| \quad (\mathbf{w}_*)\| = \left\|\mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}[\quad (\ \mathbf{w}_* \ \mathbf{x} \quad)]\right\| \leq \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}[\| \quad (\ \mathbf{w}_* \ \mathbf{x} \quad)\|] \leq$$

Next we study the case

$$\frac{1}{\ \ 2} \quad \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\| \leq 2$$

Fro 9 we have

$$
\begin{aligned}
& (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) + \frac{}{2}\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|^2 \\
\leq\ & (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\widehat{\mathbf{w}}) \quad \widehat{\ }(\mathbf{w}_*)] \ \widehat{\mathbf{w}} \quad \mathbf{w}_* \ + \quad (\mathbf{w}_*) \quad \widehat{\ }(\mathbf{w}_*) \ \widehat{\mathbf{w}} \quad \mathbf{w}_* \\
=\ & (\widehat{\mathbf{w}}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\widehat{\mathbf{w}}) \quad \widehat{\ }(\mathbf{w}_*)] \ \widehat{\mathbf{w}} \quad \mathbf{w}_* \ + \quad (\mathbf{w}_*) \quad \widehat{\ }(\mathbf{w}_*) \ \widehat{\mathbf{w}} \quad \mathbf{w}_* \\
\leq\ & \underbrace{\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\|\widehat{\mathbf{w}}-\mathbf{w}_*\|} \left\langle \quad (\mathbf{w}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\mathbf{w}) \quad \widehat{\ }(\mathbf{w}_*)] \ \mathbf{w} \quad \mathbf{w}_* \right\rangle}_{:=B_1} \\
& + \underbrace{\left\| \quad (\mathbf{w}_*) \quad \widehat{\ }(\mathbf{w}_*)\right\|}_{:=B_2} \|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|
\end{aligned}
$$

e rst bound $_1$ o ut ze the fact the rando var ab e $\|\widehat{\mathbf{w}} \quad \mathbf{w}_*\|$ es n the range (1 ^{2} 2] we deve op the fo ow ng e a

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

$$\frac{1}{\ \ 2} \quad \leq 2$$

the following bound holds:

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma} \left\langle \quad (\mathbf{w}) \quad (\mathbf{w}_*) \quad [\ \widehat{\ }(\mathbf{w}) \quad \widehat{\ }(\mathbf{w}_*)] \ \mathbf{w} \quad \mathbf{w}_* \right\rangle \leq \frac{4 \quad ^2}{\ \ } \left(8 + \sqrt{2\log -}\right)$$

where $= 2\log_2(\ ) + \log_2(2 \ )$.

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

$$_1 < \frac{4 \, \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\; - \;} \left(8 + \sqrt{2 \log -} \right) = \frac{\| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\; - \;} \qquad (44)$$

where s de ned n

e then proceed to hand e $_2$ 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 , *we have*

$$\left\| \;\; (\mathbf{w}_*) \;\;\; \widehat{\;}(\mathbf{w}_*) \right\| < \frac{2 \;\; \log(2 \;)}{\;} + \sqrt{\frac{8 \;\; _* \log(2 \;)}{\;}} \qquad (4)$$

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

$$\begin{aligned} (\widehat{\mathbf{w}}) \;\; (\mathbf{w}_*) + \frac{\;}{2} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2 \\ < \frac{\| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\; - \;} + \frac{2 \;\; \log(2 \;) \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|}{\;} + \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \| \sqrt{\frac{8 \;\; _* \log(2 \;)}{\;}} \end{aligned} \qquad (4)$$

e subst tute

$$\frac{\| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\; - \;} < \frac{2 \;\; 2 \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\;} + \frac{\;}{4} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2$$

$$\| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \| \sqrt{\frac{8 \;\; _* \log(2 \;)}{\;}} < \frac{8 \;\; _* \log(2 \;)}{\;} + \frac{\;}{4} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2$$

nto (4) and then have

$$\begin{aligned} (\widehat{\mathbf{w}}) \;\; (\mathbf{w}_*) &< \frac{2 \;\; 2 \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\;} + \frac{2 \;\; \log(2 \;) \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|}{\;} + \frac{8 \;\; _* \log(2 \;)}{\;} \\ &\overset{(8)}{<} \frac{4 \;\; 2 \;\; 2 \;\; 2}{\;} + \frac{4 \;\;\; \log(2 \;)}{\;} + \frac{8 \;\; _* \log(2 \;)}{\;} \end{aligned}$$

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

o prove we subst tute

$$\frac{2 \;\; \log(2 \;) \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|}{\;} < \frac{8 \;\; ^2 \log^2(2 \;)}{2} + \frac{\;}{8} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2$$

$$\| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \| \sqrt{\frac{8 \;\; _* \log(2 \;)}{\;}} < \frac{16 \;\; _* \log(2 \;)}{\;} + \frac{\;}{8} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2$$

nto (4) and then have

$$\begin{aligned} (\widehat{\mathbf{w}}) \;\; (\mathbf{w}_*) + \frac{\;}{4} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2 \\ < \frac{\| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2}{\; - \;} + \frac{8 \;\; ^2 \log^2(2 \;)}{2} + \frac{16 \;\; _* \log(2 \;)}{\;} \\ < \frac{\;}{4} \| \widehat{\mathbf{w}} \;\; \mathbf{w}_* \|^2 + \frac{8 \;\; ^2 \log^2(2 \;)}{2} + \frac{16 \;\; _* \log(2 \;)}{\;} \end{aligned}$$

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 $(1 \quad ^2 \; 2 \;]$ nto $= 2\log_2(\;) + \log_2(2 \;)$ consecut ve seg ents $\Delta_1 \; \Delta_2 \quad \Delta_s$ such that

$$\Delta_k = \Bigg(\underbrace{\frac{2^{k-1}}{2}}_{:=\gamma_k^-} \; \underbrace{\frac{2^k}{2}}_{:=\gamma_k^+} \Bigg] \qquad = 1$$

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

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma} \Big\langle \quad (\mathbf{w}) \qquad (\mathbf{w}_*) \quad [\quad \widehat{\;}(\mathbf{w}) \qquad \widehat{\;}(\mathbf{w}_*)] \; \mathbf{w} \quad \mathbf{w}_* \Big\rangle$$
$$\le \sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+} \Big\langle \quad (\mathbf{w}) \qquad (\mathbf{w}_*) \quad [\quad \widehat{\;}(\mathbf{w}) \qquad \widehat{\;}(\mathbf{w}_*)] \; \mathbf{w} \quad \mathbf{w}_* \Big\rangle \qquad (47)$$

Based on the McD ar d s nequa ty McD ar d 9 89 and the ade acher co p ex ty Bart ett and Mende son 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 , *we have*

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+} \Big\langle \quad (\mathbf{w}) \qquad (\mathbf{w}_*) \quad [\quad \widehat{\;}(\mathbf{w}) \qquad \widehat{\;}(\mathbf{w}_*)] \; \mathbf{w} \quad \mathbf{w}_* \Big\rangle$$
$$\le \frac{(\; _k^+)^2}{\sqrt{\;}} \left(8 + \sqrt{2\log\frac{1}{\;}} \right) \qquad (48)$$

nce $\quad \Delta_k$ we have

$$_k^+ = 2 \; _k^- \le 2 \qquad (49)$$

hus w th probab ty at east 1 we have

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma} \Big\langle \quad (\mathbf{w}) \qquad (\mathbf{w}_*) \quad [\quad \widehat{\;}(\mathbf{w}) \qquad \widehat{\;}(\mathbf{w}_*)] \; \mathbf{w} \quad \mathbf{w}_* \Big\rangle$$
$$\overset{(47),(48),(49)}{\le} \frac{4 \quad ^2}{\sqrt{\;}} \left(8 + \sqrt{2\log\frac{1}{\;}} \right)$$

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

Appendix G. Proof of Lemma 7

o s p fy the notat on we de ne

$$_i(\mathbf{w}) = (\; \mathbf{w} \; \mathbf{x}_i \quad _i) \qquad = 1$$

$$(\; _1 \qquad _n) = \sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+} \Big\langle \quad (\mathbf{w}) \qquad (\mathbf{w}_*) \quad \frac{1}{\;}\sum_{i=1}^{n} [\quad _i(\mathbf{w}) \qquad _i(\mathbf{w}_*)] \; \mathbf{w} \quad \mathbf{w}_* \Big\rangle$$

o upper bound $(\; _1 \qquad _n)$ we ut ze the McD ar d s nequa ty McD ar d 9 89

Theorem 8 *Let* $_1$ $_n$ *be independent random variables taking values in a set* A*, and assume that* $: A^n$ $\mathbb{R}$ *satisfies*

$$\sup_{x_1,\ldots,x_n,x_i'\in A} \Big| \ (\ _1 \qquad _n) \qquad (\ _1 \qquad _{i-1} \ \ '_i \ \ _{i+1} \qquad _n)\Big| \leq \ _i$$

for every $1 \leq \ \leq$ *. Then, for every* 0*,*

$$(\ _1 \qquad _n) \quad \mathrm{E}[\ (\ _1 \qquad _n)] \quad \leq \exp\left(\frac{2 \ ^2}{\sum_{i=1}^n \ ^2_i} \right)$$

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

$$\ _i(\mathbf{w}) \qquad _i(\mathbf{w}_*) \ \mathbf{w} \quad \mathbf{w}_* \ \leq \ \ \mathbf{w} \quad \mathbf{w}_* \ ^2 \leq \ (\ ^+_k)^2$$

As a resu t when a rando funct on $_i$ changes the rando var ab e $(\ _1 \qquad _n)$ can change by no ore than $2 \ (\ ^+_k)^2$ o see th s we have

$$(\ _1 \qquad _n) \quad (\ _1 \qquad _{i-1} \ \ '_i \ \ _{i+1} \qquad _n)$$

$$\leq \frac{1}{} \sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \Big\langle \ '_i(\mathbf{w}) \qquad '_i(\mathbf{w}_*) \quad [\ _i(\mathbf{w}) \qquad _i(\mathbf{w}_*)] \ \mathbf{w} \quad \mathbf{w}_* \Big\rangle \leq \frac{2}{} \ (\ ^+_k)^2$$

McD ar d s nequa ty p es that w th probab ty at east 1

$$(\ _1 \qquad _n) \leq \mathrm{E}[\ (\ _1 \qquad _n)] + \ (\ ^+_k)^2 \sqrt{\frac{2}{} \log \frac{1}{}}$$

Let $(\ '_1 \qquad '_n)$ be an ndependent copy of $(\ _1 \qquad _n)$ and $_1$ $_n$ be 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 $\mathrm{E}[\ (\ _1 \qquad _n)]$ as fo ows

$$\mathrm{E}_{h_1,\ldots,h_n}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \left\langle \quad (\mathbf{w}) \qquad (\mathbf{w}_*) \quad \frac{1}{}\sum_{i=1}^n [\ _i(\mathbf{w}) \qquad _i(\mathbf{w}_*)] \ \mathbf{w} \quad \mathbf{w}_* \right\rangle \right]$$

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

$$\left. \mathrm{E}_{h_1',\ldots,h_n'}\left[\sum_{i=1}^n \Big\langle \ '_i(\mathbf{w}) \qquad '_i(\mathbf{w}_*) \ \mathbf{w} \quad \mathbf{w}_* \Big\rangle \right] \quad \sum_{i=1}^n \ _i(\mathbf{w}) \qquad _i(\mathbf{w}_*) \ \mathbf{w} \quad \mathbf{w}_* \right]$$

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

$$\left. \sum_{i=1}^n \Big\langle \ '_i(\mathbf{w}) \qquad '_i(\mathbf{w}_*) \ \mathbf{w} \quad \mathbf{w}_* \Big\rangle \quad \sum_{i=1}^n \ _i(\mathbf{w}) \qquad _i(\mathbf{w}_*) \ \mathbf{w} \quad \mathbf{w}_* \right]$$

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

$$
\sum_{i}^{n}
$$

Note that ^{2} s 2 L psch tz over [ ] and $_i(\mathbf{w}) + {}_i(\mathbf{w})$ [2 ^+_k , 2 ^+_k] 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 we have

$$
\begin{aligned}
&\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i(\ {}_i(\mathbf{w}) + \ {}_i(\mathbf{w}))^2\right]\\
&\le 4\ {}^+_k\ \sqrt{\ }\ \mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i(\ {}_i(\mathbf{w}) + \ {}_i(\mathbf{w}))\right]\\
&\le 4\ {}^+_k\ \sqrt{\ }\left(\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ {}_i(\mathbf{w})\right] + \mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ {}_i(\mathbf{w})\right]\right)
\end{aligned}
$$

ar y we have

$$
\begin{aligned}
&\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i(\ {}_i(\mathbf{w})\quad {}_i(\mathbf{w}))^2\right]\\
&\le 4\ {}^+_k\ \sqrt{\ }\left(\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ {}_i(\mathbf{w})\right] + \mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ {}_i(\mathbf{w})\right]\right) \qquad 4
\end{aligned}
$$

Co b n ng and 4 we arr ve at

$$
\begin{aligned}
&\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ \ {}_i(\mathbf{w})\quad {}_i(\mathbf{w}_*)\ \mathbf{w}\quad \mathbf{w}_*\right]\\
&\le 2\ {}^+_k\ \sqrt{\ }\left(\underbrace{\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ {}_i(\mathbf{w})\right]}_{:=C_1} + \underbrace{\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+}\sum_{i=1}^n {}_i\ {}_i(\mathbf{x})\right]}_{:=C_2}\right)
\end{aligned}
$$

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

$$
\begin{aligned}
|\ {}_i(\mathbf{w})\quad {}_i(\mathbf{w}')| &= \frac{1}{\ }|\ {}'(\ \mathbf{w}\ \mathbf{x}_i\quad {}_i)\quad {}'(\ \mathbf{w}'\ \mathbf{x}_i\quad {}_i)|\\
&\le \sqrt{\ }\ |\ \mathbf{w}\ \mathbf{x}_i\quad \mathbf{w}'\ \mathbf{x}_i\ | = \sqrt{\ }\ |\ \mathbf{x}_i\ \mathbf{w}\quad \mathbf{w}_*\quad \mathbf{x}_i\ \mathbf{w}'\quad \mathbf{w}_*\ |
\end{aligned}
$$

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

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