We sincerely thank all reviewers for your interest in the paper and the insightful reviews!

[Responses to R#1] **Q1. Can we estimate the bounds from data?** Yes! This is an important question, and we will add a new corollary stating that all our bounds can be computed from data. In fact, obtaining a preliminary estimate on $\mathcal{L}^*$ is a step within our two-stage procedure for the variance-dependent rate (see lines 167-172). A very similar analysis answers this question: we can simply use $\widehat{\mathcal{L}^*} := \mathbb{P}_n \ell(\hat{h}_{\text{ERM}}; z)$ to estimate $\mathcal{L}^*$, and use $\widehat{\mathcal{V}^*} := \mathbb{P}_n[\ell(\hat{h}_{\text{ERM}}; z)^2] - \mathbb{P}_n \ell(\hat{h}_{\text{ERM}}; z)$ to estimate $\mathcal{V}^*$. Furthermore, we can reuse the samples for pure evaluation purpose. Similar to the inequality above line 181, we can bound both $(\widehat{\mathcal{L}^*} - \mathcal{L}^*)^2$ and $|\widehat{V} - \mathcal{V}^*|$ by $O(r^*)$. This precision is enough to rewrite our original bounds by $\widehat{\mathcal{L}^*}$ and $\widehat{\mathcal{V}^*}$, with other quantities unchanged in order. **Q2. Do we require more efforts to find ψ?** We first note that all previous analyses also require knowing ψ because they rely on knowledge of r^*—the fixed point of $B\psi$ (see lines 162-166). When one know the covering numbers, one standard choice of ψ is Dudley's integral used in Examples 1,2. These illustrate that there is little additional efforts in this identification. **Q3. ψ is dependent on n?** Yes, by definition one must take different ψ for different n (e.g., when taking $n \to \infty$, ψ should be 0). The standard notion "Rademacher complexity" also depends on fixed n. **Notation.** We will clarify that $a \vee b = \max\{a, b\}$ in the preliminaries.

[Response to R#2] **Q2. Can the bounds be described for neural networks?** Yes, though the description contains eigenvalues that are hard to compute analytically. Our theory *systematically* provides improved problem-dependent rates as long as one can find good ψ and the class is rich. A line of recent works on infinitely wide neural networks consider the equivalence between the prediction function found by gradient descent, and the RKHS induced by the "Neural Tangent Kernel." Many of these works explicitly express the resulting kernel matrix, so our theorems are applicable as illustrated in Example 3. However, our bounds contain eigenvalues of the kernel matrix, and it is difficult to assess their decay pattern without further analysis. **Q2. Explain why traditional analysis is optimal for parametric classes?** We will add the following explanation under line 100. Due to the conceptual proof (2.6), the gap between our result and the traditional analysis originates from the "sub-root" inequality $\psi(r;\delta)/\sqrt{r} \le \psi(r^*;\delta)/\sqrt{r^*}$, which is true for all sub-root ψ. This inequality becomes an equality when $\psi(r;\delta) = O(\sqrt{dr/n})$ in the parametric case. However, when $\mathcal{F}$ is rich, $\psi(r;\delta)/\sqrt{r}$ will be strictly decreasing so that the "sub-root" inequality can be loose (e.g., in Example 1, $\psi = O(\sqrt{r^{1-\rho}/n})$ so that $\sqrt{\psi(r;\delta)}/\sqrt{r} = O(\sqrt{1/(n \cdot r^{\rho})})$. The richer $\mathcal{F}$ is, the more improvement from our theory. **Writing.** We will reorganize Sections 2-3, and do our best to make the comparison in Section 6 clearer.

[Responses to R#3] We are glad to see your appreciation of our machinery! We hope our techniques can become standard tools to prove adaptive generalization error bounds. **Q1. Trade-off between optimality and practicality?** There is indeed a trade-off between statistical performance and computation. Similar to majority of previous works [12, 18, 5], our moment-penalized estimator does not preserve convexity of the population risk, while ERM and the estimator in [16] do preserve that convexity. In our answer to R#1's Q1, we explain how to compute the bounds from data. When choosing among different estimators, one can estimate different bounds to decide whether the added price of optimization results in suitable gains to make it worthwhile. **Q2. Optimality of our results?** A short answer is that, both our variance-dependent rate and loss-dependent rate exhibit optimal *direct dependence on* n when the excess loss class satisfies standard metric entropy growth conditions. For example, when $\log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H} - \ell \circ h^*; L_2(\mathbb{P}_n)) \le O(\varepsilon^{-2\rho})$ for a fixed $\rho \in (0, 1)$, both rates match the optimal direct dependence on n given by Dudley's integral. Judging whether the variance-dependent rate is optimal in all regimes requires constructing a particular class of problems where $\text{Var}[\ell(h^*; z)] = \mathcal{V}^*$. Although we strongly believe it can be done under a suitable minimax framework, we do not have a rigorous proof yet. Our loss-dependent rate is proposed for the particular algorithm ERM so the minimax framework requires further restrictions.

[Responses to R#4] Thank you for your throughout reading! The typo list is very helpful, and we will carefully check the whole manuscript. There are indeed typos on the B factor, but all our rates are actually sharp on B. The *generic correction* is: our loss/variance-dependent rates are $\psi(\mathcal{V}^*;\delta) \vee \frac{r^*}{B} \vee \frac{B\log(1/\delta)}{n}$ and $\psi(B\mathcal{L}^*;\delta) \vee \frac{r^*}{B} \vee \frac{B\log(1/\delta)}{n}$; the previously best known loss/variance-dependent rates are $\sqrt{\mathcal{L}^* r^*/B} \vee \frac{r^*}{B} \vee \frac{B\log(1/\delta)}{n}$ and $\sqrt{\mathcal{V}^* r^*/B^2} \vee \frac{r^*}{B} \vee \frac{B\log(1/\delta)}{n}$. **Q1. $\mathbb{P}f$ in inequality (2.6)?** We agree that it is better and clearer to firstly write $\mathbb{P}f$ in the last term of (2.6), then explains that $\mathbb{P}f$ is close to $\mathbb{P}_n f$ when evaluated at a fixed f, and finally contrast this term to the result of the traditional analysis. **Q2. Comparison in line 144?** In line 207 we explain that for most classes of interests, r^* will be at least of order $\frac{B^2}{n}$ (this is the order of r^* for a one-dimensional class). We will explain this before line 144 so that we only need to compare the orders of $B\mathcal{L}^*$ and r^*. **On Theorem 2 and line 209.** In the result of Theorem 2 we should correct r^* to r^*/B, and line 209 is correct (see our *generic correction*). Indeed, as r^* is the fixed point of $B\psi$, whenever one want to take it outside ψ, the order should be $\frac{r^*}{B}$. **Correction to VC classes.** That term should be corrected to $\log(B^2/\mathcal{V}^*)$, and the regime in which we improve all known results is actually $B^2/(\log n)^{\alpha} \le \mathcal{V}^*$ with arbitrary fixed $\alpha > 0$. Still, this is the first result that closes the notorious $O(\log n)$ gap without invoking any further assumptions on $\mathcal{H}$ (e.g., the complicated "capacity function" assumption in [5]). However, as richer classes exhibit much more improvements, we will shorten the discussion on VC classes and expand the discussion on kernel classes.