A. Proof of Proposition 1

Proof. Let e_j's denote standard basis vectors. We have

$$\nabla_s \phi_{\mathrm{LN}}(s,y) = -\sum_{j=1}^{m} P_j(y) e_j + \sum_{j=1}^{m} \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} e_j$$

Therefore,

$$\begin{aligned}\|\nabla_s \phi_{\mathrm{LN}}(s,y)\|_1 &\le \sum_{j=1}^{m} P_j(y)\|e_j\|_1 + \sum_{j=1}^{m} \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})}\|e_j\|_1 \\ &= 2.\end{aligned}$$

We also have

$$[\nabla_s^2 \phi_{\mathrm{LN}}(s,y)]_{j,k} = \begin{cases} -\frac{\exp(2s_j)}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} + \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} & \text{if } j = k \\ -\frac{\exp(s_j+s_k)}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} & \text{if } j \neq k\ . \end{cases}$$

Moreover,

$$\begin{aligned}\|\nabla_s^2 \phi_{\mathrm{LN}}(s,y)\|_{\infty\to 1} &\le \sum_{j=1}^{m}\sum_{k=1}^{m} |[\nabla_s^2 \phi_{\mathrm{LN}}(s,y)]_{j,k}| \\ &\le \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{\exp(s_j+s_k)}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} + \sum_{j=1}^{m} \frac{\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} \\ &= \frac{(\sum_{j=1}^{m}\exp(s_j))^2}{(\sum_{j'=1}^{m}\exp(s_{j'}))^2} + \frac{\sum_{j=1}^{m}\exp(s_j)}{\sum_{j'=1}^{m}\exp(s_{j'})} \\ &= 2\end{aligned}$$

□

B. Proof of Proposition 2

Proof. Let $1_{(\text{condition})}$ denote an indicator variable. We have

$$[\nabla_s \phi_{\mathrm{SD}}(s,y)]_j = D(1)\left(\sum_{i=1}^{m} G(r_i)\left[\frac{1}{\sigma}\frac{\exp(s_i/\sigma)}{\sum_{j'}\exp(s_{j'}/\sigma)} 1_{(i=j)} - \frac{1}{\sigma}\frac{\exp((s_i+s_j)/\sigma)}{(\sum_{j'}\exp(s_{j'}/\sigma))^2}\right]\right)$$

Therefore,

$$\begin{aligned}\frac{\|\nabla_s \phi_{\mathrm{SD}}(s,y)\|_1}{D(1)G(Y_{\max})} &\le \sum_{j=1}^{m}\left(\sum_{i=1}^{m}\left[\frac{1}{\sigma}\frac{\exp(s_i/\sigma)}{\sum_{j'}\exp(s_{j'}/\sigma)} 1_{(i=j)} + \frac{1}{\sigma}\frac{\exp((s_i+s_j)/\sigma)}{(\sum_{j'}\exp(s_{j'}/\sigma))^2}\right]\right) \\ &= \frac{1}{\sigma}\left(\frac{\sum_j \exp(s_j/\sigma)}{\sum_{j'}\exp(s_{j'}/\sigma)} + \frac{(\sum_j \exp(s_j/\sigma))^2}{(\sum_{j'}\exp(s_{j'}/\sigma))^2}\right) \\ &= \frac{2}{\sigma}.\end{aligned}$$

□

C. RankSVM

The RankSVM surrogate is defined as:

$$\phi_{RS}(s,y) = \sum_{i=1}^{m}\sum_{j=1}^{m} \max(0, 1_{(y_i > y_j)}(1 + s_j - s_i))$$

It is easy to see that $\nabla_s \phi_{RS}(s,y) = \sum_{i=1}^m \sum_{j=1}^m \max(0, 1_{(y_i > y_j)}(1 + s_j - s_i))(e_j - e_i)$. Thus, the ℓ_1 norm of gradient is $O(m^2)$.

D. Proof of Theorem 3

Proof. It is straightforward to check that $\mathcal{F}'_{\text{lin}}$ is contained in both $\mathcal{F}_{\text{full}}$ as well as $\mathcal{F}_{\text{perminv}}$. So, we just need to prove that any f that is in both $\mathcal{F}_{\text{full}}$ and $\mathcal{F}_{\text{perminv}}$ has to be in $\mathcal{F}'_{\text{lin}}$ as well.

Let P_π denote the $m \times m$ permutation matrix corresponding to a permutation π. Consider the full linear class $\mathcal{F}_{\text{full}}$. In matrix notation, the permutation invariance property means that, for any π, X, we have $P_\pi[\langle X, W_1 \rangle, \ldots, \langle X, W_m \rangle\rangle]^\top = [\langle P_\pi X, W_1 \rangle, \ldots, \langle P_\pi X, W_m \rangle]^\top$.

Let $\rho_1 = \{P_\pi : \pi(1) = 1\}$, where $\pi(i)$ denotes the index of the element in the ith position according to permutation π. Fix any $P \in \rho_1$. Then, for any X, $\langle X, W_1 \rangle = \langle PX, W_1 \rangle$. This implies that, for all X, $\text{Tr}({W_1}^\top X) = \text{Tr}({W_1}^\top PX)$. Using the fact that $\text{Tr}(A^\top X) = \text{Tr}(B^\top X), \forall X$ implies $A = B$, we have that ${W_1}^\top = {W_1}^\top P$. Because $P^\top = P^{-1}$, this means $PW_1 = W_1$. This shows that all rows of W_1, other than 1st row, are the same but perhaps different from 1st row. By considering $\rho_i = \{P_\pi : \pi(i) = i\}$ for $i > 1$, the same reasoning shows that, for each i, all rows of W_i, other than ith row, are the same but possibly different from ith row.

Let $\rho_{1\leftrightarrow 2} = \{P_\pi : \pi(1) = 2, \pi(2) = 1\}$. Fix any $P \in \rho_{1\leftrightarrow 2}$. Then, for any X, $\langle X, W_2 \rangle = \langle PX, W_1 \rangle$ and $\langle X, W_1 \rangle = \langle PX, W_2 \rangle$. Thus, we have $W_2^\top = W_1^\top P$ as well as $W_1^\top = W_2^\top P$ which means $PW_2 = W_1, PW_1 = W_2$. This shows that row 1 of W_1 and row 2 of W_2 are the same. Moreover, row 2 of W_1 and row 1 of W_2 are the same. Thus, for some $u, u' \in \mathbb{R}^d$, W_1 is of the form $[u|u'|u'|\ldots|u']^\top$ and W_2 is of the form $[u'|u|u'|\ldots|u']^\top$. Repeating this argument by considering $\rho_{1\leftrightarrow i}$ for $i > 2$ shows that W_i is of the same form (u in row i and u' elsewhere).

Therefore, we have proved that any linear map that is permutation invariant has to be of the form:

$$X \mapsto \left(u^\top X_i + (u')^\top \sum_{j \neq i} X_j \right)_{i=1}^m .$$

We can reparameterize above using $w = u - u'$ and $v = u'$ which proves the result. □

E. Proof of Lemma 4

Proof. The first equality is true because

$$\begin{aligned} \|X^\top\|_{1\to p} &= \sup_{v \neq 0} \frac{\|X^\top v\|_p}{\|v\|_1} = \sup_{v \neq 0} \sup_{u \neq 0} \frac{\langle X^\top v, u \rangle}{\|v\|_1 \|u\|_q} \\ &= \sup_{u \neq 0} \sup_{v \neq 0} \frac{\langle v, Xu \rangle}{\|v\|_1 \|u\|_q} = \sup_{u \neq 0} \frac{\|Xu\|_\infty}{\|u\|_q} = \|X\|_{q \to \infty}. \end{aligned}$$

The second is true because

$$\begin{aligned} \|X\|_{q\to\infty} &= \sup_{u \neq 0} \frac{\|Xu\|_\infty}{\|u\|_q} = \sup_{u \neq 0} \max_{j=1}^m \frac{|\langle X_j, u \rangle|}{\|u\|_q} \\ &= \max_{j=1}^m \sup_{u \neq 0} \frac{|\langle X_j, u \rangle|}{\|u\|_q} = \max_{j=1}^m \|X_j\|_p. \end{aligned}$$

□

F. Proof of Theorem 6

Our theorem is developed from the "expectation version" of Theorem 6 of Shalev-Shwartz et al. (2009) that was originally given in probabilistic form. The expected version is as follows.

Let $\mathcal{Z}$ be a space endowed with a probability distribution generating iid draws $Z_1, \ldots, Z_n$. Let $\mathcal{W} \subseteq \mathbb{R}^d$ and $f : \mathcal{W} \times \mathcal{Z} \to$

$\mathbb{R}$ be λ-strongly convex[4] and G-Lipschitz (w.r.t. $\|\cdot\|_2$) in w for every z. We define $F(w) = \mathbb{E}\left[f(w,Z)\right]$ and let

$$w^\star = \underset{w\in\mathcal{W}}{\text{argmin}}\; F(w),$$
$$\hat{w} = \underset{w\in\mathcal{W}}{\text{argmin}}\; \frac{1}{n}\sum_{i=1}^{n} f(w, Z_i).$$

Then $\mathbb{E}\left[F(\hat{w}) - F(w^\star)\right] \le \frac{4G^2}{\lambda n}$, where the expectation is taken over the sample. The above inequality can be proved by carefully going through the proof of Theorem 6 proved by Shalev-Shwartz et al. (2009).

We now derive the "expectation version" of Theorem 7 of Shalev-Shwartz et al. (2009). Define the regularized empirical risk minimizer as follows:

$$\hat{w}_\lambda = \underset{w\in\mathcal{W}}{\text{argmin}}\; \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} f(w, Z_i). \tag{9}$$

The following result gives optimality guarantees for the regularized empirical risk minimizer.

Theorem 18. *Let $\mathcal{W} = \{w : \|w\|_2 \le W_2\}$ and let $f(w,z)$ be convex and G-Lipschitz (w.r.t. $\|\cdot\|_2$) in w for every z. Let $Z_1, ..., Z_n$ be iid samples and let $\lambda = \sqrt{\frac{\frac{4G^2}{n}}{\frac{W_2^2}{2} + \frac{4W_2^2}{n}}}$. Then for $\hat{w}_\lambda$ and $w^\star$ as defined above, we have*

$$\mathbb{E}\left[F(\hat{w}_\lambda) - F(w^\star)\right] \le 2\,G\,W_2\left(\frac{8}{n} + \sqrt{\frac{2}{n}}\right). \tag{10}$$

Proof. Let $r_\lambda(w,z) = \frac{\lambda}{2}\|w\|_2^2 + f(w,z)$. Then r_λ is λ-strongly convex with Lipschitz constant $\lambda W_2 + G$ in $\|\cdot\|_2$. Applying "expectation version" of Theorem 6 of Shalev-Shwartz et al. (2009) to r_λ, we get

$$\mathbb{E}\left[\frac{\lambda}{2}\|\hat{w}_\lambda\|_2^2 + F(\hat{w}_\lambda)\right] \le \min_{w\in\mathcal{W}}\left\{\frac{\lambda}{2}\|w\|_2^2 + F(w)\right\} + \frac{4(\lambda W_2 + G)^2}{\lambda n} \le \frac{\lambda}{2}\|w^\star\|_2^2 + F(w^*) + \frac{4(\lambda W_2 + G)^2}{\lambda n}.$$

Thus, we get

$$\mathbb{E}\left[F(\hat{w}_\lambda) - F(w^\star)\right] \le \frac{\lambda W_2^2}{2} + \frac{4(\lambda W_2 + G)^2}{\lambda n}.$$

Minimizing the upper bound w.r.t. λ, we get $\lambda = \sqrt{\frac{4G^2}{n}}\sqrt{\frac{1}{\frac{W_2^2}{2} + \frac{4W_2^2}{n}}}$. Plugging this choice back in the equation above and using the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ finishes the proof of Theorem 18. □

We now have all ingredients to prove Theorem 6.

Proof of Theorem 6. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $f(w,z) = \phi(Xw, y)$ and apply Theorem 18. Finally note that if ϕ is G_ϕ-Lipschitz w.r.t. $\|\cdot\|_\infty$ and every row of $X \in \mathbb{R}^{m\times d}$ has Euclidean norm bounded by R_X then $f(\cdot, z)$ is $G_\phi R_X$-Lipschitz w.r.t. $\|\cdot\|_2$ in w. □

G. Proof of Theorem 12

Proof. Following exactly the same line of reasoning (reducing a sample of size n, where each prediction is $\mathbb{R}^m$-valued, to an sample of size mn, where each prediction is real valued) as in the beginning of proof of Proposition 7, we have

$$\mathcal{N}_\infty(\epsilon, \phi\circ\mathcal{F}_1, n) \le \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_1, mn). \tag{11}$$

Plugging in the following bound due to Zhang (2002, Corollary 5):

$$\log_2 \mathcal{N}_\infty(\epsilon/G_\phi, \mathcal{G}_1, mn) \le \left\lceil \frac{288\,G_\phi^2\,W_1^2\,\bar{R}_X^2\,(2 + \ln d)}{\epsilon^2} \right\rceil \times \log_2\left(2\lceil 8G_\phi W_1 \bar{R}_X/\epsilon\rceil mn + 1\right)$$

into (11) respectively proves the result. □

[4] Recall that a function is called λ-strongly convex (w.r.t. $\|\cdot\|_2$) iff $f - \frac{\lambda}{2}\|\cdot\|_2^2$ is convex.

H. Calculations involved in deriving Equation (8)

Plugging in the value of η from (7) into the expression

$$\frac{L_\phi(w^\star)}{(1-4\eta H)} + \frac{W_2^2}{2\eta(1-4\eta H)n}$$

yields (using the shorthand $L^\star$ for $L_\phi(w^\star)$)

$$L^\star + \frac{2HW_2L^\star}{\sqrt{4H^2W_2^2+2HL^\star n}} + \frac{W_2}{n}\left[\frac{4H^2W_2^2}{\sqrt{4H^2W_2^2+2HL^\star n}} + \sqrt{4H^2W_2^2+2HL^\star n} + 4HW_2\right]$$

Denoting HW_2^2/n by x, this simplifies to

$$L^\star + \frac{2\sqrt{x}L^\star + 4x\sqrt{x}}{\sqrt{4x+2L^\star}} + \sqrt{x}\sqrt{4x+2L^\star} + 4x.$$

Using the arithmetic mean-geometric mean inequality to upper bound the middle two terms gives

$$L^\star + 2\sqrt{2xL^\star + 4x^2} + 4x.$$

Finally, using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we get our final upper bound

$$L^\star + 2\sqrt{2xL^\star} + 8x.$$

I. Calculation of smoothness constant

$$\begin{aligned}
&\|(X^{(i)})^\top \nabla_s^2 \phi(X^{(i)}w, y^{(i)})X^{(i)}\|_{2\to 2} = \sup_{v\neq 0} \frac{\|(X^{(i)})^\top \nabla_s^2 \phi(X^{(i)}w, y^{(i)})X^{(i)}v\|_2}{\|v\|_2} \\
&\le \sup_{v\neq 0} \frac{\|(X^{(i)})^\top\|_{1\to 2}\|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})X^{(i)}v\|_1}{\|v\|_2} \le \sup_{v\neq 0} \frac{\|(X^{(i)})^\top\|_{1\to 2}\cdot\|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1}\cdot\|X^{(i)}v\|_\infty}{\|v\|_2} \\
&\le \sup_{v\neq 0} \frac{\|(X^{(i)})^\top\|_{1\to 2}\cdot\|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1}\cdot\|X^{(i)}\|_{2\to\infty}\cdot\|v\|_2}{\|v\|_2} \\
&\le \left(\max_{j=1}^{m}\|X_j^{(i)}\|\right)^2 \cdot \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1} \\
&\le R_X^2 \|\nabla_s^2 \phi(X^{(i)}w, y^{(i)})\|_{\infty\to 1}.
\end{aligned}$$

J. Proof of Lemma 14

Proof. Consider the function

$$f(t) = \phi((1-t)s_1 + ts_2).$$

It is clearly non-negative. Moreover

$$\begin{aligned}
|f'(t_1) - f'(t_2)| &= |\langle \nabla_s\phi(s_1 + t_1(s_2-s_1)) - \nabla_s\phi(s_1 + t_2(s_2-s_1)), s_2 - s_1\rangle| \\
&\le |||\nabla_s\phi(s_1 + t_1(s_2-s_1)) - \nabla_s\phi(s_1 + t_2(s_2-s_1))|||_\star \cdot |||s_2 - s_1||| \\
&\le H_\phi\,|t_1 - t_2|\,|||s_2 - s_1|||^2
\end{aligned}$$

and therefore it is smooth with constant $h = H_\phi|||s_2 - s_1|||^2$. Appealing to Lemma 13 now gives

$$(f(1) - f(0))^2 \le 6H_\phi|||s_2 - s_1|||^2(f(1) + f(0))(1-0)^2$$

which proves the lemma since $f(0) = \phi(s_1)$ and $f(1) = \phi(s_2)$. □

K. Proof of Proposition 15

Proof. Let $w, w' \in \mathcal{F}_{\phi,2}(r)$. Using Lemma 14

$$
\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)}w, y^{(i)}) - \phi(X^{(i)}w', y^{(i)})\right)^2 \\
&\le 6H_\phi \sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)}w, y^{(i)}) + \phi(X^{(i)}w', y^{(i)})\right) \\
&\quad \cdot \|X^{(i)}w - X^{(i)}w'\|_\infty^2 \\
&\le 6H_\phi \cdot \max_{i=1}^{n} \|X^{(i)}w - X^{(i)}w'\|_\infty^2 \\
&\quad \cdot \sum_{i=1}^{n} \frac{1}{n} \left(\phi(X^{(i)}w, y^{(i)}) + \phi(X^{(i)}w', y^{(i)})\right) \\
&= 6H_\phi \cdot \max_{i=1}^{n} \|X^{(i)}w - X^{(i)}w'\|_\infty^2 \cdot \left(\hat{L}_\phi(w) + \hat{L}_\phi(w')\right) \\
&\le 12H_\phi r \cdot \max_{i=1}^{n} \|X^{(i)}w - X^{(i)}w'\|_\infty^2 .
\end{aligned}
$$

where the last inequality follows because $\hat{L}_\phi(w) + \hat{L}_\phi(w') \le 2r$.
This immediately implies that if we have a cover of the class $\mathcal{G}_2$ at scale $\epsilon/\sqrt{12H_\phi r}$ w.r.t. the metric

$$
\max_{i=1}^{n} \max_{j=1}^{m} \left| \left\langle X_j^{(i)}, w \right\rangle - \left\langle X_j^{(i)}, w' \right\rangle \right|
$$

then it is also a cover of $\mathcal{F}_{\phi,2}(r)$ w.r.t. $d_2^{Z^{(1:n)}}$. Therefore, we have

$$
\mathcal{N}_2(\epsilon, \mathcal{F}_{\phi,2}(r), Z^{(1:n)}) \le \mathcal{N}_\infty(\epsilon/\sqrt{12H_\phi r}, \mathcal{G}_2, mn). \tag{12}
$$

Appealing once again to a result by Zhang (2002, Corollary 3), we get

$$
\log_2 \mathcal{N}_\infty(\epsilon/\sqrt{12H_\phi r}, \mathcal{G}_2, mn) \le \left\lceil \frac{12H_\phi\, W_2^2\, R_X^2\, r}{\epsilon^2} \right\rceil \times \log_2(2mn+1)
$$

which finishes the proof. □

L. Proof of Corollary 16

Proof. We plug in Proposition 15's estimate into (5):

$$
\begin{aligned}
\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{\phi,2}(r)\right) &\le \inf_{\alpha>0} \left(4\alpha + 10 \int_\alpha^{\sqrt{Br}} \sqrt{\frac{\left\lceil \frac{12H_\phi\, W_2^2\, R_X^2\, r}{\epsilon^2} \right\rceil \log_2(2mn+1)}{n}} d\epsilon \right) \\
&\le \inf_{\alpha>0} \left(4\alpha + 20\sqrt{3} W_2 R_X \sqrt{\frac{r H_\phi \log_2(3mn)}{n}} \int_\alpha^{\sqrt{Br}} \frac{1}{\epsilon} d\epsilon \right) .
\end{aligned}
$$

Now choosing $\alpha = C\sqrt{r}$ where $C = 5\sqrt{3} W_2 R_X \sqrt{\frac{H_\phi \log_2(3mn)}{n}}$ gives us the upper bound

$$
\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{\phi,2}(r)\right) \le 4\sqrt{r} C \left(1 + \log \frac{\sqrt{B}}{C}\right) \le 4\sqrt{r} C \log \frac{3\sqrt{B}}{C}.
$$

□

M. Proof of Theorem 17

Proof. We appeal to Theorem 6.1 of Bousquet (2002) that assumes there exists an upper bound

$$\widehat{\mathfrak{R}}_n\left(\mathcal{F}_{2,\phi}(r)\right) \le \psi_n(r)$$

where $\psi_n : [0,\infty) \to \mathbb{R}_+$ is a non-negative, non-decreasing, non-zero function such that $\psi_n(r)/\sqrt{r}$ is non-increasing. The upper bound in Corollary 16 above satisfies these conditions and therefore we set $\psi_n(r) = 4\sqrt{r}C\log\frac{3\sqrt{B}}{C}$ with C as defined in Corollary 16. From Bousquet's result, we know that, with probability at least $1-\delta$,

$$\forall w \in \mathcal{F}_2,\ L_\phi(w) \le \hat{L}_\phi(w) + 45r_n^\star + \sqrt{8r_n^\star L_\phi(w)} + \sqrt{4r_0 L_\phi(w)} + 20r_0$$

where $r_0 = B(\log(1/\delta) + \log\log n)/n$ and $r_n^\star$ is the largest solution to the equation $r = \psi_n(r)$. In our case, $r_n^\star = \left(4C\log\frac{3\sqrt{B}}{C}\right)^2$. This proves the first inequality.

Now, using the above inequality with $w = \hat{w}$, the empirical risk minimizer and noting that $\hat{L}_\phi(\hat{w}) \le \hat{L}_\phi(w^\star)$, we get

$$L_\phi(\hat{w}) \le \hat{L}_\phi(w^\star) + 45r_n^\star + \sqrt{8r_n^\star L_\phi(\hat{w})} + \sqrt{4r_0 L_\phi(\hat{w})} + 20r_0$$

The second inequality now follows after some elementary calculations detailed below. □

M.1. Details of some calculations in the proof of Theorem 17

Using Bernstein's inequality, we have, with probability at least $1-\delta$,

$$\begin{aligned}
\hat{L}_\phi(w^\star) &\le L_\phi(w^\star) + \sqrt{\frac{4\mathrm{Var}[\phi(Xw^\star, y)]\log(1/\delta)}{n}} + \frac{4B\log(1/\delta)}{n} \\
&\le L_\phi(w^\star) + \sqrt{\frac{4BL_\phi(w^\star)\log(1/\delta)}{n}} + \frac{4B\log(1/\delta)}{n} \\
&\le L_\phi(w^\star) + \sqrt{4r_0 L_\phi(w^\star)} + 4r_0.
\end{aligned}$$

Set $D_0 = 45r_n^\star + 20r_0$. Putting the two bounds together and using some simple upper bounds, we have, with probability at least $1-2\delta$,

$$\begin{aligned}
L_\phi(\hat{w}) &\le \sqrt{D_0\hat{L}_\phi(w^\star)} + D_0, \\
\hat{L}_\phi(w^\star) &\le \sqrt{D_0 L_\phi(w^\star)} + D_0.
\end{aligned}$$

which implies that

$$L_\phi(\hat{w}) \le \sqrt{D_0}\sqrt{\sqrt{D_0 L_\phi(w^\star)} + D_0} + D_0.$$

Using $\sqrt{ab} \le (a+b)/2$ to simplify the first term on the right gives us

$$L_\phi(\hat{w}) \le \frac{D_0}{2} + \frac{\sqrt{D_0 L_\phi(w^\star)} + D_0}{2} + D_0 = \frac{\sqrt{D_0 L_\phi(w^\star)}}{2} + 2D_0\ .$$