We relate sequential risk to statistical risk, assuming the data sequence on which an online algorithm is run is generated by independent and identically distributed draws from a fixed and unknown distribution $\mathcal{D}$.

Fix a convex and differentiable loss function ℓ, for example $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ for regression or $\ell(\widehat{y}, y) = [1 - y\widehat{y}]_+$ for classification. As usual with online learning, we focus on linear predictors $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. The statistical risk of a linear predictor $\boldsymbol{w} \in \mathbb{R}^d$ is defined by

$$\ell_{\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}\Big[\ell\big(\boldsymbol{w}^\top \boldsymbol{X}, Y\big)\Big]$$

where $(\boldsymbol{X}, Y)$ is drawn from $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$.

In statistical learning a training set S is a random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_m, Y_m)$ drawn from $\mathcal{D}$. This induces a sequence $\ell_1, \ldots, \ell_m$ of convex loss functions defined by $\ell_t(\boldsymbol{w}) = \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_t, y_t\big)$. When running an online learning algorithm, such as OGD, on this sequence we obtain a corresponding sequence $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ of linear predictors. We want to derive an upper bound on the statistical risk of a linear predictor derived from this sequence. In particular, we consider the *average predictor*

$$\overline{\boldsymbol{w}} = \frac{1}{m} \sum_{t=1}^{m} \boldsymbol{w}_t \ .$$

Since ℓ is convex in $\boldsymbol{w}$, Jensen inequality gives us

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) = \mathbb{E}\Big[\ell\big(\overline{\boldsymbol{w}}^\top \boldsymbol{X}, Y\big)\Big] \leq \mathbb{E}\left[\frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}, Y\big)\right] = \frac{1}{m} \sum_{t=1}^{m} \ell_{\mathcal{D}}(\boldsymbol{w}_t)$$

where the last equality holds because of linearity of expectation. Hence, the risk of the average predictor is upper bounded by the average risk of the predictors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$.

The crucial step is to connect the average statistical risk to the sequential risk. Observe that, under the assumption that S is a statistical sample, $\boldsymbol{w}_t$ is determined by the first $t - 1$ examples $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{t-1}, Y_{t-1})$. Therefore, by applying the definition of risk to the expected value of the loss of $\boldsymbol{w}_t$ on the t-th example $(\boldsymbol{X}_t, Y_t)$, we can write

$$\mathbb{E}\Big[\ell_{\mathcal{D}}(\boldsymbol{w}_t) - \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t\big) \,\Big|\, (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{t-1}, Y_{t-1})\Big] = 0 \ . \tag{1}$$

The above equality means the following: if we condition on the first $t - 1$ examples, then $\boldsymbol{w}_t$ is determined, and the expected value of $\ell_t(\boldsymbol{w}_t)$ with respect to the draw on the t-th example is —by definition— the risk of $\boldsymbol{w}_t$.

We write $\mathbb{E}_{t-1}$ to denote expectation conditioned on $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{t-1}, Y_{t-1})$. If we sum both sides of (1) over $t = 1, \ldots, m$ and divide by m we get

$$\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_{t-1} \Big[\ell_{\mathcal{D}}(\boldsymbol{w}_t) - \ell\big(\boldsymbol{w}_t^{\top} \boldsymbol{X}_t, Y_t\big) \Big] = 0 \ .$$

For each $t = 1, \ldots, m$ let Z_t be the random variable $\ell_{\mathcal{D}}(\boldsymbol{w}_t) - \ell\big(\boldsymbol{w}_t^{\top} \boldsymbol{X}_t, Y_t\big)$. Then $Z_1, \ldots, Z_m$ are all functions of the same random sample S, and such that

$$\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_{t-1}[Z_t] = 0 \ .$$

We assume $\ell_t \in [0, M]$ so that $|Z_t| \leq M$. Bounded random variables $Z_1, Z_2, \ldots$ such that $\mathbb{E}_{t-1}[Z_t] = 0$ are called *martingale difference sequence* with increments bounded by M. Although these random variables are not independent, we can still prove a Chernoff-Hoeffding bound of the form

$$\frac{1}{m} \sum_{t=1}^{m} Z_t \leq 2M \sqrt{\frac{1}{2m} \ln \frac{1}{\delta}}$$

with probability at least $1 - \delta$ with respect to the random draw of S. This implies

$$\frac{1}{m} \sum_{t=1}^{m} \ell_{\mathcal{D}}(\boldsymbol{w}_t) \leq \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^{\top} \boldsymbol{X}_t, Y_t\big) + M \sqrt{\frac{2}{m} \ln \frac{1}{\delta}} \tag{2}$$

again with probability at least $1 - \delta$. As for the average predictor $\overline{\boldsymbol{w}}$, the result we obtain can be formulated as

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) \leq \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^{\top} \boldsymbol{x}_t, y_t\big) + \mathcal{O}\left(\frac{1}{\sqrt{m}}\right) \qquad \text{with high probability.}$$

In other words, the statistical risk of the average predictor is bounded in probability by the sequential risk on the training set.

We can work a bit more to obtain a risk bound based on the analysis of the sequential risk. Consider for example regression with quadratic loss. If we run OGD with projection onto the set $\big\{ \boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| \leq U \big\}$, and assume $\|\boldsymbol{x}_t\| \leq X$ and $|y_t| \leq UX$ for each t, we get that for *every* realization $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ of the training set

$$\frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^{\top} \boldsymbol{x}_t, y_t\big) \leq \min_{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| \leq U} \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{u}^{\top} \boldsymbol{x}_t, y_t\big) + 8(UX)^2 \sqrt{\frac{2}{m}} \ .$$

Substituting the right-hand side in (2), and observing that for the square loss $M = 4(UX)^2$, we can then write

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) \leq \min_{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| \leq U} \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{u}^{\top} \boldsymbol{X}_t, Y_t\big) + 12(UX)^2 \sqrt{\frac{2}{m} \ln \frac{2}{\delta}}$$

with probability at least $1 - \delta/2$ with respect to the random draw of S.

Finally, letting

$$\boldsymbol{u}^* = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d:\|\boldsymbol{u}\|\leq U} \ell_{\mathcal{D}}(\boldsymbol{u})$$

we clearly have

$$\min_{\boldsymbol{u}\in\mathbb{R}^d:\|\boldsymbol{u}\|\leq U} \frac{1}{m}\sum_{t=1}^{m}\ell\big(\boldsymbol{u}^\top\boldsymbol{x}_t,y_t\big) \leq \frac{1}{m}\sum_{t=1}^{m}\ell\big(\boldsymbol{x}_t^\top\boldsymbol{u}^*,y_t\big) \ .$$

Since, for each $t = 1, \ldots, m$ we have $\mathbb{E}\big[\ell\big(\boldsymbol{X}_t^\top\boldsymbol{u}^*,Y_t\big)\big] = \ell_{\mathcal{D}}(\boldsymbol{u}^*)$, we can apply the standard Chernoff-Hoeffding bond and derive

$$\frac{1}{m}\sum_{t=1}^{m}\ell\big(\boldsymbol{X}_t^\top\boldsymbol{u}^*,Y_t\big) \leq \ell_{\mathcal{D}}(\boldsymbol{u}^*) + 4(UX)^2\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}} \qquad \text{with probability at least } 1 - \delta/2.$$

We then got the following explicit bound on the risk of the average predictor

$$\ell_{\mathcal{D}}(\overline{\boldsymbol{w}}) \leq \ell_{\mathcal{D}}(\boldsymbol{u}^*) + 14(UX)^2\sqrt{\frac{2}{m}\ln\frac{2}{\delta}}$$

with probability at least $1 - \delta$ with respect to the random draw of S.