

CSE 541: Interactive Learning

Lecture 4: Online squared loss, mixability, and online convex optimization

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Monday, October 12, 2026



Squared loss

Fast rates, in the batch and online worlds

Regression with a finite class

- $(X, Y) \sim \nu$ on $\mathcal{X} \times [0, 1]$; a finite class $\mathcal{F}$ of predictors $f : \mathcal{X} \rightarrow [0, 1]$.
- Risk and empirical risk: $L(f) = \mathbb{E}[(f(X) - Y)^2]$, $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$.
- ERM $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{L}_n(f)$; best in class $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$.

Regression with a finite class

- $(X, Y) \sim \nu$ on $\mathcal{X} \times [0, 1]$; a finite class $\mathcal{F}$ of predictors $f : \mathcal{X} \rightarrow [0, 1]$.
- Risk and empirical risk: $L(f) = \mathbb{E}[(f(X) - Y)^2]$, $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$.
- ERM $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{L}_n(f)$; best in class $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$.

Bayes predictor

$f^*(x) = \mathbb{E}[Y \mid X = x]$ minimizes L over *all* functions, and for any f ,
 $L(f) = L(f^*) + \mathbb{E}[(f(X) - f^*(X))^2]$ (the cross term vanishes by the tower property). In general $f^* \notin \mathcal{F}$.

Regression with a finite class

- $(X, Y) \sim \nu$ on $\mathcal{X} \times [0, 1]$; a finite class $\mathcal{F}$ of predictors $f : \mathcal{X} \rightarrow [0, 1]$.
- Risk and empirical risk: $L(f) = \mathbb{E}[(f(X) - Y)^2]$, $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$.
- ERM $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{L}_n(f)$; best in class $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$.

Bayes predictor

$f^*(x) = \mathbb{E}[Y \mid X = x]$ minimizes L over *all* functions, and for any f ,
 $L(f) = L(f^*) + \mathbb{E}[(f(X) - f^*(X))^2]$ (the cross term vanishes by the tower property). In general $f^* \notin \mathcal{F}$.

Slow rate. $(f(x) - y)^2 \in [0, 1]$, so Hoeffding plus the union bound gives, exactly as for classification,
 $L(\hat{f}) - L(f_{\mathcal{F}}^*) \leq \sqrt{2 \log(2|\mathcal{F}|/\delta)/n}$ w.p. $\geq 1 - \delta$. Only boundedness was used.

Regression with a finite class

- $(X, Y) \sim \nu$ on $\mathcal{X} \times [0, 1]$; a finite class $\mathcal{F}$ of predictors $f : \mathcal{X} \rightarrow [0, 1]$.
- Risk and empirical risk: $L(f) = \mathbb{E}[(f(X) - Y)^2]$, $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$.
- ERM $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{L}_n(f)$; best in class $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$.

Bayes predictor

$f^*(x) = \mathbb{E}[Y \mid X = x]$ minimizes L over *all* functions, and for any f ,
 $L(f) = L(f^*) + \mathbb{E}[(f(X) - f^*(X))^2]$ (the cross term vanishes by the tower property). In general $f^* \notin \mathcal{F}$.

Slow rate. $(f(x) - y)^2 \in [0, 1]$, so Hoeffding plus the union bound gives, exactly as for classification, $L(\hat{f}) - L(f_{\mathcal{F}}^*) \leq \sqrt{2 \log(2|\mathcal{F}|/\delta)/n}$ w.p. $\geq 1 - \delta$. Only boundedness was used.

Classification had a **fast rate** $\log |\mathcal{H}|/n$ when realizable; is there an analogue here? We need a concentration inequality that sees the *variance*.

The loss difference W^f

For $f \in \mathcal{F}$ let $W^f(x, y) = (f(x) - y)^2 - (f^*(x) - y)^2$, the **loss difference** against the Bayes predictor.

Lemma (realizable case)

Suppose $f^*(x) = \mathbb{E}[Y \mid X = x]$ belongs to $\mathcal{F}$. Then for every $f \in \mathcal{F}$: (1) $|W^f| \leq 1$. (2) $\mathbb{E}[W^f(X, Y)] = L(f) - L(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$. (3) $\text{Var}[W^f(X, Y)] \leq 4(L(f) - L(f^*))$.

The loss difference W^f

For $f \in \mathcal{F}$ let $W^f(x, y) = (f(x) - y)^2 - (f^*(x) - y)^2$, the **loss difference** against the Bayes predictor.

Lemma (realizable case)

Suppose $f^*(x) = \mathbb{E}[Y \mid X = x]$ belongs to $\mathcal{F}$. Then for every $f \in \mathcal{F}$: (1) $|W^f| \leq 1$. (2) $\mathbb{E}[W^f(X, Y)] = L(f) - L(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$. (3) $\text{Var}[W^f(X, Y)] \leq 4(L(f) - L(f^*))$.

(1) Both squared losses lie in $[0, 1]$, so their difference lies in $[-1, 1]$.

The loss difference W^f

For $f \in \mathcal{F}$ let $W^f(x, y) = (f(x) - y)^2 - (f^*(x) - y)^2$, the **loss difference** against the Bayes predictor.

Lemma (realizable case)

Suppose $f^*(x) = \mathbb{E}[Y \mid X = x]$ belongs to $\mathcal{F}$. Then for every $f \in \mathcal{F}$: (1) $|W^f| \leq 1$. (2) $\mathbb{E}[W^f(X, Y)] = L(f) - L(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$. (3) $\text{Var}[W^f(X, Y)] \leq 4(L(f) - L(f^*))$.

(1) Both squared losses lie in $[0, 1]$, so their difference lies in $[-1, 1]$.

(2) $\mathbb{E}[W^f] = L(f) - L(f^*)$ by definition of L . For the second equality write $f - Y = (f - f^*) + (f^* - Y)$ and expand:

$$L(f) = \mathbb{E}[(f - Y)^2] =$$

The loss difference W^f

For $f \in \mathcal{F}$ let $W^f(x, y) = (f(x) - y)^2 - (f^*(x) - y)^2$, the **loss difference** against the Bayes predictor.

Lemma (realizable case)

Suppose $f^*(x) = \mathbb{E}[Y \mid X = x]$ belongs to $\mathcal{F}$. Then for every $f \in \mathcal{F}$: (1) $|W^f| \leq 1$. (2) $\mathbb{E}[W^f(X, Y)] = L(f) - L(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$. (3) $\text{Var}[W^f(X, Y)] \leq 4(L(f) - L(f^*))$.

(1) Both squared losses lie in $[0, 1]$, so their difference lies in $[-1, 1]$.

(2) $\mathbb{E}[W^f] = L(f) - L(f^*)$ by definition of L . For the second equality write $f - Y = (f - f^*) + (f^* - Y)$ and expand:

$$L(f) = \mathbb{E}[(f - Y)^2] = \mathbb{E}[(f - f^*)^2] + 2\mathbb{E}[(f - f^*)(f^* - Y)] + L(f^*),$$

and the cross term is $\mathbb{E}[(f - f^*) \mathbb{E}[f^* - Y \mid X]] = 0$ by the tower property, **because** $f^* = \mathbb{E}[Y \mid X]$ (for another $g \in \mathcal{F}$ in place of f^* it would not vanish).

The loss difference W^f

For $f \in \mathcal{F}$ let $W^f(x, y) = (f(x) - y)^2 - (f^*(x) - y)^2$, the **loss difference** against the Bayes predictor.

Lemma (realizable case)

Suppose $f^*(x) = \mathbb{E}[Y \mid X = x]$ belongs to $\mathcal{F}$. Then for every $f \in \mathcal{F}$: (1) $|W^f| \leq 1$. (2) $\mathbb{E}[W^f(X, Y)] = L(f) - L(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$. (3) $\text{Var}[W^f(X, Y)] \leq 4(L(f) - L(f^*))$.

(1) Both squared losses lie in $[0, 1]$, so their difference lies in $[-1, 1]$.

(2) $\mathbb{E}[W^f] = L(f) - L(f^*)$ by definition of L . For the second equality write $f - Y = (f - f^*) + (f^* - Y)$ and expand:

$$L(f) = \mathbb{E}[(f - Y)^2] = \mathbb{E}[(f - f^*)^2] + 2\mathbb{E}[(f - f^*)(f^* - Y)] + L(f^*),$$

and the cross term is $\mathbb{E}[(f - f^*) \mathbb{E}[f^* - Y \mid X]] = 0$ by the tower property, **because $f^* = \mathbb{E}[Y \mid X]$** (for another $g \in \mathcal{F}$ in place of f^* it would not vanish).

(3) Factor $W^f = (f - f^*)(f + f^* - 2y)$; since $f, f^*, y \in [0, 1]$, $|f + f^* - 2y| \leq 2$, so

$$\text{Var}[W^f] \leq \mathbb{E}[(W^f)^2] = \mathbb{E}[(f - f^*)^2 (f + f^* - 2Y)^2] \leq 4\mathbb{E}[(f - f^*)^2] = 4(L(f) - L(f^*)). \blacksquare$$

Lemma (Bernstein's inequality)

Let $X_1, \dots, X_n$ be independent with $|X_i| \leq B$ and $\frac{1}{n} \sum_i \text{Var}(X_i) \leq \sigma^2$. With probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2B \log(2/\delta)}{3n}.$$

Bernstein's inequality, applied to W^f

Lemma (Bernstein's inequality)

Let $X_1, \dots, X_n$ be independent with $|X_i| \leq B$ and $\frac{1}{n} \sum_i \text{Var}(X_i) \leq \sigma^2$. With probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2B \log(2/\delta)}{3n}.$$

Apply it to $X_i = W^f(x_i, y_i)$ with $B = 1$ and $\sigma^2 = 4(L(f) - L(f^*))$ from the lemma; note $\frac{1}{n} \sum_i W^f(x_i, y_i) = \hat{L}_n(f) - \hat{L}_n(f^*)$. Union bound over $f \in \mathcal{F}$, $\beta = \log(2|\mathcal{F}|/\delta)$: w.p. $\geq 1 - \delta$, for all $f \in \mathcal{F}$,

$$(L(f) - L(f^*)) - (\hat{L}_n(f) - \hat{L}_n(f^*)) \leq \sqrt{\frac{8(L(f) - L(f^*))\beta}{n}} + \frac{2\beta}{3n}.$$

Bernstein's inequality, applied to W^f

Lemma (Bernstein's inequality)

Let $X_1, \dots, X_n$ be independent with $|X_i| \leq B$ and $\frac{1}{n} \sum_i \text{Var}(X_i) \leq \sigma^2$. With probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2B \log(2/\delta)}{3n}.$$

Apply it to $X_i = W^f(x_i, y_i)$ with $B = 1$ and $\sigma^2 = 4(L(f) - L(f^*))$ from the lemma; note $\frac{1}{n} \sum_i W^f(x_i, y_i) = \hat{L}_n(f) - \hat{L}_n(f^*)$. Union bound over $f \in \mathcal{F}$, $\beta = \log(2|\mathcal{F}|/\delta)$: w.p. $\geq 1 - \delta$, for all $f \in \mathcal{F}$,

$$(L(f) - L(f^*)) - (\hat{L}_n(f) - \hat{L}_n(f^*)) \leq \sqrt{\frac{8(L(f) - L(f^*))\beta}{n}} + \frac{2\beta}{3n}.$$

Read it

Hoeffding's $(b - a)^2/4$ is replaced by the true variance, so the gap between f 's empirical and true excess risk is $\sqrt{(\text{true excess risk})/n}$ rather than $\sqrt{1/n}$: functions close to f^* concentrate faster.

Theorem (realizable ERM, squared loss)

If $f^* = \mathbb{E}[Y \mid X] \in \mathcal{F}$, then with probability at least $1 - \delta$, $L(\hat{f}) - L(f^*) \leq \frac{28 \log(2|\mathcal{F}|/\delta)}{3n}$.

Theorem (realizable ERM, squared loss)

If $f^* = \mathbb{E}[Y | X] \in \mathcal{F}$, then with probability at least $1 - \delta$, $L(\hat{f}) - L(f^*) \leq \frac{28 \log(2|\mathcal{F}|/\delta)}{3n}$.

Proof. Work on the Bernstein event of the previous slide and write $\Delta = L(\hat{f}) - L(f^*) \geq 0$.

Step 1: where realizability is used. $\hat{f}$ minimizes $\hat{L}_n$ over $\mathcal{F}$ and $f^* \in \mathcal{F}$, so $\hat{L}_n(\hat{f}) \leq \hat{L}_n(f^*)$, i.e. $\hat{L}_n(\hat{f}) - \hat{L}_n(f^*) \leq 0$.

Fast rate for ERM with squared loss

Theorem (realizable ERM, squared loss)

If $f^* = \mathbb{E}[Y | X] \in \mathcal{F}$, then with probability at least $1 - \delta$, $L(\hat{f}) - L(f^*) \leq \frac{28 \log(2|\mathcal{F}|/\delta)}{3n}$.

Proof. Work on the Bernstein event of the previous slide and write $\Delta = L(\hat{f}) - L(f^*) \geq 0$.

Step 1: where realizability is used. $\hat{f}$ minimizes $\hat{L}_n$ over $\mathcal{F}$ and $f^* \in \mathcal{F}$, so $\hat{L}_n(\hat{f}) \leq \hat{L}_n(f^*)$, i.e. $\hat{L}_n(\hat{f}) - \hat{L}_n(f^*) \leq 0$.

Step 2: plug $f = \hat{f}$ into the uniform bound. The empirical term is ≤ 0 , so

$$\Delta \leq$$

Fast rate for ERM with squared loss

Theorem (realizable ERM, squared loss)

If $f^* = \mathbb{E}[Y | X] \in \mathcal{F}$, then with probability at least $1 - \delta$, $L(\hat{f}) - L(f^*) \leq \frac{28 \log(2|\mathcal{F}|/\delta)}{3n}$.

Proof. Work on the Bernstein event of the previous slide and write $\Delta = L(\hat{f}) - L(f^*) \geq 0$.

Step 1: where realizability is used. $\hat{f}$ minimizes $\hat{L}_n$ over $\mathcal{F}$ and $f^* \in \mathcal{F}$, so $\hat{L}_n(\hat{f}) \leq \hat{L}_n(f^*)$, i.e. $\hat{L}_n(\hat{f}) - \hat{L}_n(f^*) \leq 0$.

Step 2: plug $f = \hat{f}$ into the uniform bound. The empirical term is ≤ 0 , so

$$\Delta \leq \sqrt{\frac{8\Delta\beta}{n}} + \frac{2\beta}{3n}.$$

Step 3: the unknown is on both sides. AM-GM, $\sqrt{xy} \leq \frac{x+y}{2}$ with $x = \Delta$ and $y = 8\beta/n$, gives $\sqrt{8\Delta\beta/n} \leq \Delta/2 + 4\beta/n$. Hence $\Delta \leq \frac{\Delta}{2} + \frac{4\beta}{n} + \frac{2\beta}{3n} = \frac{\Delta}{2} + \frac{14\beta}{3n}$, i.e. $\Delta \leq \frac{28\beta}{3n}$. ■

Theorem (realizable ERM, squared loss)

If $f^* = \mathbb{E}[Y | X] \in \mathcal{F}$, then with probability at least $1 - \delta$, $L(\hat{f}) - L(f^*) \leq \frac{28 \log(2|\mathcal{F}|/\delta)}{3n}$.

Proof. Work on the Bernstein event of the previous slide and write $\Delta = L(\hat{f}) - L(f^*) \geq 0$.

Step 1: where realizability is used. $\hat{f}$ minimizes $\hat{L}_n$ over $\mathcal{F}$ and $f^* \in \mathcal{F}$, so $\hat{L}_n(\hat{f}) \leq \hat{L}_n(f^*)$, i.e. $\hat{L}_n(\hat{f}) - \hat{L}_n(f^*) \leq 0$.

Step 2: plug $f = \hat{f}$ into the uniform bound. The empirical term is ≤ 0 , so

$$\Delta \leq \sqrt{\frac{8\Delta\beta}{n}} + \frac{2\beta}{3n}.$$

Step 3: the unknown is on both sides. AM-GM, $\sqrt{xy} \leq \frac{x+y}{2}$ with $x = \Delta$ and $y = 8\beta/n$, gives $\sqrt{8\Delta\beta/n} \leq \Delta/2 + 4\beta/n$. Hence $\Delta \leq \frac{\Delta}{2} + \frac{4\beta}{n} + \frac{2\beta}{3n} = \frac{\Delta}{2} + \frac{14\beta}{3n}$, i.e. $\Delta \leq \frac{28\beta}{3n}$. ■

Realizability was used only in Step 1 (and in the lemma); nothing was assumed about the noise level $L(f^*)$.

Convexity, not realizability: the geometry

Now drop $f^* \in \mathcal{F}$; the natural comparator is $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$. Write $\|g\|^2 = \mathbb{E}[g(X)^2]$ and $\langle g, h \rangle = \mathbb{E}[g(X)h(X)]$.

Step 1: $f_{\mathcal{F}}^*$ is a projection. The Bayes decomposition says $L(g) = L(f^*) + \|g - f^*\|^2$ for every g , so minimizing L over $\mathcal{F}$ is minimizing the L^2 distance to f^* : $f_{\mathcal{F}}^*$ is the L^2 -projection of f^* onto $\mathcal{F}$.

Convexity, not realizability: the geometry

Now drop $f^* \in \mathcal{F}$; the natural comparator is $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$. Write $\|g\|^2 = \mathbb{E}[g(X)^2]$ and $\langle g, h \rangle = \mathbb{E}[g(X)h(X)]$.

Step 1: $f_{\mathcal{F}}^*$ is a projection. The Bayes decomposition says $L(g) = L(f^*) + \|g - f^*\|^2$ for every g , so minimizing L over $\mathcal{F}$ is minimizing the L^2 distance to f^* : $f_{\mathcal{F}}^*$ is the L^2 -projection of f^* onto $\mathcal{F}$.

Step 2: excess risk against $f_{\mathcal{F}}^*$. For $f \in \mathcal{F}$, expand $f - f^* = (f - f_{\mathcal{F}}^*) + (f_{\mathcal{F}}^* - f^*)$:

$$L(f) - L(f_{\mathcal{F}}^*) = \|f - f^*\|^2 - \|f_{\mathcal{F}}^* - f^*\|^2 = \|f - f_{\mathcal{F}}^*\|^2 + 2 \langle f - f_{\mathcal{F}}^*, f_{\mathcal{F}}^* - f^* \rangle.$$

Realizable case: $f_{\mathcal{F}}^* = f^*$, the cross term is 0, and this is the identity $L(f) - L(f^*) = \|f - f^*\|^2$ from the lemma.

Convexity, not realizability: the geometry

Now drop $f^* \in \mathcal{F}$; the natural comparator is $f_{\mathcal{F}}^* \in \arg \min_{f \in \mathcal{F}} L(f)$. Write $\|g\|^2 = \mathbb{E}[g(X)^2]$ and $\langle g, h \rangle = \mathbb{E}[g(X)h(X)]$.

Step 1: $f_{\mathcal{F}}^*$ is a projection. The Bayes decomposition says $L(g) = L(f^*) + \|g - f^*\|^2$ for every g , so minimizing L over $\mathcal{F}$ is minimizing the L^2 distance to f^* : $f_{\mathcal{F}}^*$ is the L^2 -projection of f^* onto $\mathcal{F}$.

Step 2: excess risk against $f_{\mathcal{F}}^*$. For $f \in \mathcal{F}$, expand $f - f^* = (f - f_{\mathcal{F}}^*) + (f_{\mathcal{F}}^* - f^*)$:

$$L(f) - L(f_{\mathcal{F}}^*) = \|f - f^*\|^2 - \|f_{\mathcal{F}}^* - f^*\|^2 = \|f - f_{\mathcal{F}}^*\|^2 + 2 \langle f - f_{\mathcal{F}}^*, f_{\mathcal{F}}^* - f^* \rangle.$$

Realizable case: $f_{\mathcal{F}}^* = f^*$, the cross term is 0, and this is the identity $L(f) - L(f^*) = \|f - f^*\|^2$ from the lemma.

Step 3: the projection inequality. If $\mathcal{F}$ is convex, then $g_{\epsilon} = (1 - \epsilon)f_{\mathcal{F}}^* + \epsilon f \in \mathcal{F}$ for $\epsilon \in (0, 1]$, and $f_{\mathcal{F}}^*$ is the minimizer, so

$$0 \leq \|g_{\epsilon} - f^*\|^2 - \|f_{\mathcal{F}}^* - f^*\|^2 = 2\epsilon \langle f - f_{\mathcal{F}}^*, f_{\mathcal{F}}^* - f^* \rangle + \epsilon^2 \|f - f_{\mathcal{F}}^*\|^2.$$

Divide by ϵ and let $\epsilon \downarrow 0$: $\langle f - f_{\mathcal{F}}^*, f_{\mathcal{F}}^* - f^* \rangle \geq 0$. Therefore $L(f) - L(f_{\mathcal{F}}^*) \geq \|f - f_{\mathcal{F}}^*\|^2$ for all $f \in \mathcal{F}$.

Convexity, not realizability: the fast rate

Step 4: the variance lemma needed only one inequality. Set $W^f = (f - Y)^2 - (f_{\mathcal{F}}^* - Y)^2$, with $f_{\mathcal{F}}^*$ in place of f^* . As before $|W^f| \leq 1$ and $\mathbb{E}[W^f] = L(f) - L(f_{\mathcal{F}}^*)$, and

$$\text{Var}[W^f] \leq \mathbb{E}[(f - f_{\mathcal{F}}^*)^2(f + f_{\mathcal{F}}^* - 2Y)^2] \leq 4\|f - f_{\mathcal{F}}^*\|^2 \leq 4(L(f) - L(f_{\mathcal{F}}^*))$$

by the projection inequality. In the realizable case the last step was an *equality* (Bayes decomposition); convexity gives it as an inequality, which is all Bernstein needs.

Convexity, not realizability: the fast rate

Step 4: the variance lemma needed only one inequality. Set $W^f = (f - Y)^2 - (f_{\mathcal{F}}^* - Y)^2$, with $f_{\mathcal{F}}^*$ in place of f^* . As before $|W^f| \leq 1$ and $\mathbb{E}[W^f] = L(f) - L(f_{\mathcal{F}}^*)$, and

$$\text{Var}[W^f] \leq \mathbb{E}[(f - f_{\mathcal{F}}^*)^2(f + f_{\mathcal{F}}^* - 2Y)^2] \leq 4\|f - f_{\mathcal{F}}^*\|^2 \leq 4(L(f) - L(f_{\mathcal{F}}^*))$$

by the projection inequality. In the realizable case the last step was an *equality* (Bayes decomposition); convexity gives it as an inequality, which is all Bernstein needs.

Step 5: rerun the proof. Step 1 used only $\widehat{L}_n(\widehat{f}) \leq \widehat{L}_n(f_{\mathcal{F}}^*)$, true since $f_{\mathcal{F}}^* \in \mathcal{F}$; Steps 2–3 used only the variance bound. What changes: a convex $\mathcal{F}$ is infinite, so the union bound is replaced by a uniform-convergence argument (Rademacher complexity) and $\log |\mathcal{F}|$ by the complexity of $\mathcal{F}$.

Convexity, not realizability: the fast rate

Step 4: the variance lemma needed only one inequality. Set $W^f = (f - Y)^2 - (f_{\mathcal{F}}^* - Y)^2$, with $f_{\mathcal{F}}^*$ in place of f^* . As before $|W^f| \leq 1$ and $\mathbb{E}[W^f] = L(f) - L(f_{\mathcal{F}}^*)$, and

$$\text{Var}[W^f] \leq \mathbb{E}[(f - f_{\mathcal{F}}^*)^2(f + f_{\mathcal{F}}^* - 2Y)^2] \leq 4\|f - f_{\mathcal{F}}^*\|^2 \leq 4(L(f) - L(f_{\mathcal{F}}^*))$$

by the projection inequality. In the realizable case the last step was an *equality* (Bayes decomposition); convexity gives it as an inequality, which is all Bernstein needs.

Step 5: rerun the proof. Step 1 used only $\widehat{L}_n(\widehat{f}) \leq \widehat{L}_n(f_{\mathcal{F}}^*)$, true since $f_{\mathcal{F}}^* \in \mathcal{F}$; Steps 2–3 used only the variance bound. What changes: a convex $\mathcal{F}$ is infinite, so the union bound is replaced by a uniform-convergence argument (Rademacher complexity) and $\log |\mathcal{F}|$ by the complexity of $\mathcal{F}$.

Consequence: realizability is the special case, improper learning is the general trick (notes §2.3.3)

- If $f^* \in \mathcal{F}$ then $f_{\mathcal{F}}^* = f^*$ and the cross term vanishes for *any* $\mathcal{F}$: no convexity needed.
- For a finite, non-convex $\mathcal{F}$, run ERM over $\text{conv}(\mathcal{F})$: an **improper** learner that outputs an average of members of $\mathcal{F}$. The hull has the same Rademacher complexity as $\mathcal{F}$, so this gives $O(\log |\mathcal{F}|/n)$ against the best $f \in \mathcal{F}$, with no realizability.

Online regression with squared loss

Online regression

Input: finite $\mathcal{F}$ of $f : \mathcal{X} \rightarrow [0, 1]$

for $t = 1, 2, \dots, T$:

x_t arrives

Learner predicts $\hat{y}_t \in [0, 1]$

$y_t \in [0, 1]$ is revealed

Learner suffers $(\hat{y}_t - y_t)^2$

Goal

$$\text{Regret}_T = \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^T (f(x_t) - y_t)^2,$$

for every sequence (x_t, y_t) .

What we already know. IID and realizable (or convex $\mathcal{F}$): ERM each round, summing the fast rate, gives $O(\log |\mathcal{F}| \cdot \log T)$ regret, as for realizable classification. Adversarial:

$(f(x_t) - y_t)^2 \in [0, 1]$, so exponential weights over $\mathcal{F}$ (sample $f_t \sim p_t$) gives $\sqrt{T \log(|\mathcal{F}|)/2}$.

Online regression with squared loss

Online regression

Input: finite $\mathcal{F}$ of $f : \mathcal{X} \rightarrow [0, 1]$

for $t = 1, 2, \dots, T$:

x_t arrives

Learner predicts $\hat{y}_t \in [0, 1]$

$y_t \in [0, 1]$ is revealed

Learner suffers $(\hat{y}_t - y_t)^2$

Goal

$$\text{Regret}_T = \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^T (f(x_t) - y_t)^2,$$

for every sequence (x_t, y_t) .

What we already know. IID and realizable (or convex $\mathcal{F}$): ERM each round, summing the fast rate, gives $O(\log |\mathcal{F}| \cdot \log T)$ regret, as for realizable classification. Adversarial:

$(f(x_t) - y_t)^2 \in [0, 1]$, so exponential weights over $\mathcal{F}$ (sample $f_t \sim p_t$) gives $\sqrt{T \log(|\mathcal{F}|)/2}$.

But nothing forces $\hat{y}_t$ to be some $f(x_t)$. We may predict any number in $[0, 1]$, e.g. the **average** of the class under p_t . This is **improper learning**: the predictor need not belong to $\mathcal{F}$.

Averaged exponential weights

Initialize: $w_1(f) = 1$ for all $f \in \mathcal{F}$

for $t = 1, 2, \dots, T$:

Predict $\hat{y}_t = \sum_f w_t(f) f(x_t) / \sum_f w_t(f)$

Observe y_t ; $w_{t+1}(f) = w_t(f) e^{-\eta(f(x_t) - y_t)^2}$

Mixability: the average beats every mixture

Theorem (averaged exponential weights)

With $\eta = \frac{1}{2}$, for every sequence $(x_1, y_1), \dots, (x_T, y_T)$ with $y_t \in [0, 1]$:

$$\text{Regret}_T = \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^T (f(x_t) - y_t)^2 \leq 2 \ln |\mathcal{F}|.$$

No $\sqrt{T}$; no dependence on T at all. Deterministic, and with no assumption on the data.

Mixability: the average beats every mixture

Theorem (averaged exponential weights)

With $\eta = \frac{1}{2}$, for every sequence $(x_1, y_1), \dots, (x_T, y_T)$ with $y_t \in [0, 1]$:

$$\text{Regret}_T = \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^T (f(x_t) - y_t)^2 \leq 2 \ln |\mathcal{F}|.$$

No $\sqrt{T}$; no dependence on T at all. Deterministic, and with no assumption on the data.

Notation for the proof. Write $\ell_t(f) = (f(x_t) - y_t)^2 \in [0, 1]$ for the loss of f on round t . The algorithm keeps

$$w_1(f) = 1, \quad w_{t+1}(f) = w_t(f) e^{-\eta \ell_t(f)}, \quad W_t = \sum_{f \in \mathcal{F}} w_t(f), \quad p_t(f) = \frac{w_t(f)}{W_t},$$

and predicts the p_t -average $\hat{y}_t = \sum_f p_t(f) f(x_t) = \mathbb{E}_{f \sim p_t}[f(x_t)]$.

Mixability: the average beats every mixture

Theorem (averaged exponential weights)

With $\eta = \frac{1}{2}$, for every sequence $(x_1, y_1), \dots, (x_T, y_T)$ with $y_t \in [0, 1]$:

$$\text{Regret}_T = \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^T (f(x_t) - y_t)^2 \leq 2 \ln |\mathcal{F}|.$$

No $\sqrt{T}$; no dependence on T at all. Deterministic, and with no assumption on the data.

Notation for the proof. Write $\ell_t(f) = (f(x_t) - y_t)^2 \in [0, 1]$ for the loss of f on round t . The algorithm keeps

$$w_1(f) = 1, \quad w_{t+1}(f) = w_t(f) e^{-\eta \ell_t(f)}, \quad W_t = \sum_{f \in \mathcal{F}} w_t(f), \quad p_t(f) = \frac{w_t(f)}{W_t},$$

and predicts the p_t -average $\hat{y}_t = \sum_f p_t(f) f(x_t) = \mathbb{E}_{f \sim p_t}[f(x_t)]$.

Strategy. Same potential W_t as for exponential weights, sandwiched between a per-round upper bound and a lower bound from the best f . The new ingredient: in the per-round step, Hoeffding's lemma is replaced by concavity.

Mixability: proof, Steps 1–2 (the potential)

Step 1: one round of the potential. Using $w_{t+1}(f) = w_t(f) e^{-\eta \ell_t(f)}$ and $p_t(f) = w_t(f)/W_t$,

$$\frac{W_{t+1}}{W_t} =$$

Mixability: proof, Steps 1–2 (the potential)

Step 1: one round of the potential. Using $w_{t+1}(f) = w_t(f) e^{-\eta \ell_t(f)}$ and $p_t(f) = w_t(f)/W_t$,

$$\frac{W_{t+1}}{W_t} = \sum_{f \in \mathcal{F}} \frac{w_t(f)}{W_t} e^{-\eta \ell_t(f)} = \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$$

Taking logs and summing over t telescopes: $\ln \frac{W_{T+1}}{W_1} = \sum_{t=1}^T \ln \frac{W_{t+1}}{W_t} = \sum_{t=1}^T \ln \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$

Mixability: proof, Steps 1–2 (the potential)

Step 1: one round of the potential. Using $w_{t+1}(f) = w_t(f) e^{-\eta \ell_t(f)}$ and $p_t(f) = w_t(f)/W_t$,

$$\frac{W_{t+1}}{W_t} = \sum_{f \in \mathcal{F}} \frac{w_t(f)}{W_t} e^{-\eta \ell_t(f)} = \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$$

Taking logs and summing over t telescopes: $\ln \frac{W_{T+1}}{W_1} = \sum_{t=1}^T \ln \frac{W_{t+1}}{W_t} = \sum_{t=1}^T \ln \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$

Step 2: lower bound by the best function. $W_1 = |\mathcal{F}|$, and a sum of positive terms is at least any one of them, so for every $f^* \in \mathcal{F}$

$$W_{T+1} \geq w_{T+1}(f^*) = \exp \left(-\eta \sum_{t=1}^T \ell_t(f^*) \right) \implies \ln \frac{W_{T+1}}{W_1} \geq -\eta \sum_{t=1}^T \ell_t(f^*) - \ln |\mathcal{F}|.$$

Mixability: proof, Steps 1–2 (the potential)

Step 1: one round of the potential. Using $w_{t+1}(f) = w_t(f) e^{-\eta \ell_t(f)}$ and $p_t(f) = w_t(f)/W_t$,

$$\frac{W_{t+1}}{W_t} = \sum_{f \in \mathcal{F}} \frac{w_t(f)}{W_t} e^{-\eta \ell_t(f)} = \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$$

Taking logs and summing over t telescopes: $\ln \frac{W_{T+1}}{W_1} = \sum_{t=1}^T \ln \frac{W_{t+1}}{W_t} = \sum_{t=1}^T \ln \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$

Step 2: lower bound by the best function. $W_1 = |\mathcal{F}|$, and a sum of positive terms is at least any one of them, so for every $f^* \in \mathcal{F}$

$$W_{T+1} \geq w_{T+1}(f^*) = \exp \left(-\eta \sum_{t=1}^T \ell_t(f^*) \right) \implies \ln \frac{W_{T+1}}{W_1} \geq -\eta \sum_{t=1}^T \ell_t(f^*) - \ln |\mathcal{F}|.$$

Putting 1 and 2 together:
$$-\eta \sum_{t=1}^T \ell_t(f^*) - \ln |\mathcal{F}| \leq \sum_{t=1}^T \ln \mathbb{E}_{f \sim p_t} \left[e^{-\eta \ell_t(f)} \right].$$

For exponential weights we now bounded each summand by Hoeffding's lemma, $\ln \mathbb{E}[e^{-\eta \ell}] \leq -\eta \mathbb{E}[\ell] + \eta^2/8$, and the $\eta^2/8$ per round is what became $\sqrt{T}$. Squared loss lets us do better.

Mixability: proof, Step 3 (no slack)

Claim. For $\eta \leq \frac{1}{2}$ and every round, $\ln \mathbb{E}_{f \sim p_t} [e^{-\eta(f(x_t) - y_t)^2}] \leq -\eta (\hat{y}_t - y_t)^2$.

Put $z_f = f(x_t) - y_t \in [-1, 1]$ and $g(z) = e^{-\eta z^2}$, so the left side is $\ln \mathbb{E}_{f \sim p_t} [g(z_f)]$.

Mixability: proof, Step 3 (no slack)

Claim. For $\eta \leq \frac{1}{2}$ and every round, $\ln \mathbb{E}_{f \sim p_t} [e^{-\eta(f(x_t) - y_t)^2}] \leq -\eta (\hat{y}_t - y_t)^2$.

Put $z_f = f(x_t) - y_t \in [-1, 1]$ and $g(z) = e^{-\eta z^2}$, so the left side is $\ln \mathbb{E}_{f \sim p_t} [g(z_f)]$.

g is concave where we need it. $g'(z) = -2\eta z e^{-\eta z^2}$ and $g''(z) = (4\eta^2 z^2 - 2\eta) e^{-\eta z^2}$, which is ≤ 0 exactly when $z^2 \leq \frac{1}{2\eta}$. For $\eta \leq \frac{1}{2}$ we have $\frac{1}{2\eta} \geq 1 \geq z_f^2$, so g is concave on $[-1, 1]$.

Mixability: proof, Step 3 (no slack)

Claim. For $\eta \leq \frac{1}{2}$ and every round, $\ln \mathbb{E}_{f \sim p_t} [e^{-\eta(f(x_t) - y_t)^2}] \leq -\eta(\hat{y}_t - y_t)^2$.

Put $z_f = f(x_t) - y_t \in [-1, 1]$ and $g(z) = e^{-\eta z^2}$, so the left side is $\ln \mathbb{E}_{f \sim p_t} [g(z_f)]$.

g is concave where we need it. $g'(z) = -2\eta z e^{-\eta z^2}$ and $g''(z) = (4\eta^2 z^2 - 2\eta) e^{-\eta z^2}$, which is ≤ 0 exactly when $z^2 \leq \frac{1}{2\eta}$. For $\eta \leq \frac{1}{2}$ we have $\frac{1}{2\eta} \geq 1 \geq z_f^2$, so g is concave on $[-1, 1]$.

Jensen, in the direction concavity allows: $\mathbb{E}[g(Z)] \leq g(\mathbb{E}[Z])$. With $Z = z_f$, $f \sim p_t$, and $\mathbb{E}_{f \sim p_t} [z_f] = \mathbb{E}_{f \sim p_t} [f(x_t)] - y_t = \hat{y}_t - y_t$,

$$\mathbb{E}_{f \sim p_t} [g(z_f)] \leq$$

Mixability: proof, Step 3 (no slack)

Claim. For $\eta \leq \frac{1}{2}$ and every round, $\ln \mathbb{E}_{f \sim p_t} [e^{-\eta(f(x_t) - y_t)^2}] \leq -\eta(\hat{y}_t - y_t)^2$.

Put $z_f = f(x_t) - y_t \in [-1, 1]$ and $g(z) = e^{-\eta z^2}$, so the left side is $\ln \mathbb{E}_{f \sim p_t} [g(z_f)]$.

g is concave where we need it. $g'(z) = -2\eta z e^{-\eta z^2}$ and $g''(z) = (4\eta^2 z^2 - 2\eta) e^{-\eta z^2}$, which is ≤ 0 exactly when $z^2 \leq \frac{1}{2\eta}$. For $\eta \leq \frac{1}{2}$ we have $\frac{1}{2\eta} \geq 1 \geq z_f^2$, so g is concave on $[-1, 1]$.

Jensen, in the direction concavity allows: $\mathbb{E}[g(Z)] \leq g(\mathbb{E}[Z])$. With $Z = z_f$, $f \sim p_t$, and $\mathbb{E}_{f \sim p_t} [z_f] = \mathbb{E}_{f \sim p_t} [f(x_t)] - y_t = \hat{y}_t - y_t$,

$$\mathbb{E}_{f \sim p_t} [g(z_f)] \leq g(\mathbb{E}_{f \sim p_t} [z_f]) = g(\hat{y}_t - y_t) = e^{-\eta(\hat{y}_t - y_t)^2}.$$

Taking logs (monotone) gives the claim. The loss of the *average* prediction is at most what the exponential mixture of losses promises: this is **mixability**.

Mixability: proof, Step 3 (no slack)

Claim. For $\eta \leq \frac{1}{2}$ and every round, $\ln \mathbb{E}_{f \sim p_t} [e^{-\eta(f(x_t) - y_t)^2}] \leq -\eta(\hat{y}_t - y_t)^2$.

Put $z_f = f(x_t) - y_t \in [-1, 1]$ and $g(z) = e^{-\eta z^2}$, so the left side is $\ln \mathbb{E}_{f \sim p_t} [g(z_f)]$.

g is concave where we need it. $g'(z) = -2\eta z e^{-\eta z^2}$ and $g''(z) = (4\eta^2 z^2 - 2\eta) e^{-\eta z^2}$, which is ≤ 0 exactly when $z^2 \leq \frac{1}{2\eta}$. For $\eta \leq \frac{1}{2}$ we have $\frac{1}{2\eta} \geq 1 \geq z_f^2$, so g is concave on $[-1, 1]$.

Jensen, in the direction concavity allows: $\mathbb{E}[g(Z)] \leq g(\mathbb{E}[Z])$. With $Z = z_f$, $f \sim p_t$, and $\mathbb{E}_{f \sim p_t} [z_f] = \mathbb{E}_{f \sim p_t} [f(x_t)] - y_t = \hat{y}_t - y_t$,

$$\mathbb{E}_{f \sim p_t} [g(z_f)] \leq g(\mathbb{E}_{f \sim p_t} [z_f]) = g(\hat{y}_t - y_t) = e^{-\eta(\hat{y}_t - y_t)^2}.$$

Taking logs (monotone) gives the claim. The loss of the *average* prediction is at most what the exponential mixture of losses promises: this is **mixability**.

Combine with Steps 1–2. Summing the claim over t :

$-\eta \sum_t \ell_t(f^*) - \ln |\mathcal{F}| \leq \sum_t \ln \mathbb{E}_{p_t} [e^{-\eta \ell_t(f)}] \leq -\eta \sum_t (\hat{y}_t - y_t)^2$. Divide by η and rearrange:
 $\sum_t (\hat{y}_t - y_t)^2 - \sum_t \ell_t(f^*) \leq \ln |\mathcal{F}| / \eta$ for every f^* ; take $\eta = \frac{1}{2}$, the largest value allowed. ■

Proper vs. improper, and the parallel with the batch world

	Classification, 0/1 loss	Regression, squared loss
Batch, slow rate	$\sqrt{\log \mathcal{H} /n}$ (Hoeffding)	$\sqrt{\log \mathcal{F} /n}$ (Hoeffding)
Batch, fast rate	$\log \mathcal{H} /n$ if realizable	$\log \mathcal{F} /n$ realizable or convex; improper : always
Regret, IID	ERM: $\log \mathcal{H} \log T$ realiz., $\sqrt{T \log \mathcal{H} }$ agnostic	ERM: $\log \mathcal{F} \log T$ realiz. or convex
Regret, online	$\sqrt{T \log \mathcal{H} /2}$, randomized, proper	$2 \ln \mathcal{F} $, deterministic, improper

Proper vs. improper, and the parallel with the batch world

	Classification, 0/1 loss	Regression, squared loss
Batch, slow rate	$\sqrt{\log \mathcal{H} /n}$ (Hoeffding)	$\sqrt{\log \mathcal{F} /n}$ (Hoeffding)
Batch, fast rate	$\log \mathcal{H} /n$ if realizable	$\log \mathcal{F} /n$ realizable or convex; improper : always
Regret, IID	ERM: $\log \mathcal{H} \log T$ realiz., $\sqrt{T \log \mathcal{H} }$ agnostic	ERM: $\log \mathcal{F} \log T$ realiz. or convex
Regret, online	$\sqrt{T \log \mathcal{H} /2}$, randomized, proper	$2 \ln \mathcal{F} $, deterministic, improper

- **Improper** = the prediction is not made by any single $f \in \mathcal{F}$. Online it buys the T -free bound (a proper learner must randomize and pays $\sqrt{T}$); in the batch world, ERM over $\text{conv}(\mathcal{F})$ removes realizability.
- Why not average for the 0/1 loss? $e^{-\eta \mathbf{1}\{\cdot\}}$ is not concave in anything useful: the 0/1 loss is *not mixable*, and any deterministic predictor can be made wrong every round.
- Curvature is the common thread: it gives the fast rate in the batch world (through the variance) and mixability in the online world. Next: classes too large to enumerate, where convexity and gradients take over.