

# CSE 541: Interactive Learning

## Lecture 3: Portfolios, squared loss, and online convex optimization

Kevin Jamieson

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

Wednesday, October 7, 2026



# Last time, in one slide

## Online learning with $d$ actions

for  $t = 1, 2, \dots, T$ : player picks  $p_t \in \Delta_d$  and plays  $I_t \sim p_t$ ;

adversary reveals  $\ell_t \in [0, 1]^d$ ; player pays  $\ell_t(I_t)$

Classification:  $d = |\mathcal{H}|$ ,  $\ell_t(i) = \mathbf{1}\{h_i(x_t) \neq y_t\}$

## Exponential weights

$w_1 \equiv 1$ ;  $p_t(i) = w_t(i)/W_t$ ,  $W_t = \sum_i w_t(i)$ ;

$w_{t+1}(i) = w_t(i) e^{-\eta \ell_t(i)}$

## Theorem

For any  $\ell_1, \dots, \ell_T \in [0, 1]^d$ ,

$\text{Regret}_T = \sum_t \langle p_t, \ell_t \rangle - \min_i \sum_t \ell_t(i) \leq \frac{\log d}{\eta} + \frac{\eta T}{8}$ ,

so  $\eta = \sqrt{8 \log d / T}$  gives  $\sqrt{T \log(d)/2}$ .

**The proof, in two lines.** Track  $\log(W_{T+1}/W_1)$ :

- *Upper:*  $\frac{W_{t+1}}{W_t} = \mathbb{E}_{I \sim p_t}[e^{-\eta \ell_t(I)}]$ , and Hoeffding's lemma gives  $\log \mathbb{E}_{I \sim p_t}[e^{-\eta \ell_t(I)}] \leq -\eta \langle p_t, \ell_t \rangle + \frac{\eta^2}{8}$ .
- *Lower:*  $W_{T+1} \geq w_{T+1}(i) = e^{-\eta \sum_t \ell_t(i)}$  for every  $i$ .

## Two messages from last time

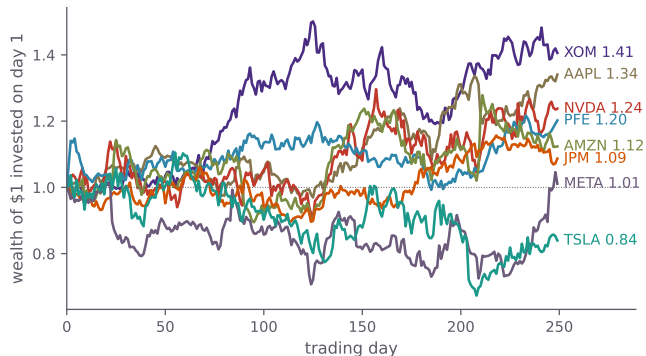
Dropping IID cost nothing: same  $\log d$ , same  $\sqrt{T}$  as ERM under IID.

The  $\eta^2/8$  slack per round is exactly what becomes  $\sqrt{T}$ . Today: losses where that slack disappears.

# Online portfolio optimization

The loss decides who you are competing with

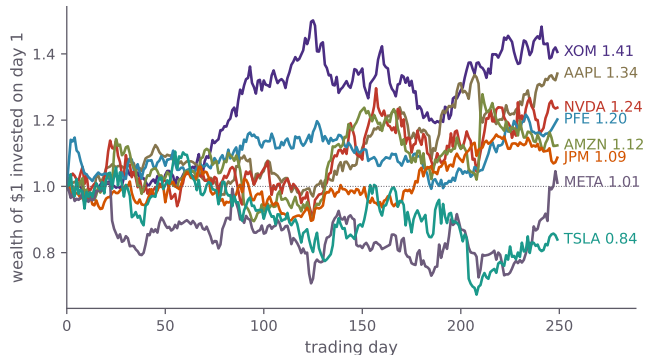
# Which stock would you have picked?



Daily closes for eight large US stocks, Sep 29, 2025 to Sep 25, 2026, each normalized to \$1 on day 1.

- The leader changes hands over and over; the year's winner was mid-pack for most of it.
- Halfway through, the eventual worst stock had been the best.

# Which stock would you have picked?



Daily closes for eight large US stocks, Sep 29, 2025 to Sep 25, 2026, each normalized to \$1 on day 1.

- The leader changes hands over and over; the year's winner was mid-pack for most of it.
- Halfway through, the eventual worst stock had been the best.

Picking the best stock is hard *even in hindsight*, let alone in advance. Can we still be competitive with it, with no model of the market at all?

# Investing without a model of the market

**Setup.**  $d$  stocks. On day  $t$ , stock  $i$  has *gross return*  $r_t(i) = \text{price}_{t+1}(i)/\text{price}_t(i) \geq 0$ .

A portfolio is  $p_t \in \Delta_d$  (wealth fractions), rebalanced daily. Starting from wealth  $v_1$ ,

$$v_{T+1} = v_1 \prod_{t=1}^T \langle p_t, r_t \rangle,$$

so the log-wealth is

$$\log(v_{T+1}/v_1) = \sum_{t=1}^T \log \langle p_t, r_t \rangle.$$

# Investing without a model of the market

**Setup.**  $d$  stocks. On day  $t$ , stock  $i$  has *gross return*  $r_t(i) = \text{price}_{t+1}(i)/\text{price}_t(i) \geq 0$ .

A portfolio is  $p_t \in \Delta_d$  (wealth fractions), rebalanced daily. Starting from wealth  $v_1$ ,

$$v_{T+1} = v_1 \prod_{t=1}^T \langle p_t, r_t \rangle,$$

so the log-wealth is

$$\log(v_{T+1}/v_1) = \sum_{t=1}^T \log \langle p_t, r_t \rangle.$$

**Classical (Markowitz):** assume  $r_t$  IID with mean  $\mu$  and covariance  $\Sigma$  estimated from history; solve  $\min_{p \in \Delta_d} p^\top \Sigma p$  s.t.  $p^\top \mu \geq \bar{r}$ . IID fails exactly when it matters.

# Investing without a model of the market

**Setup.**  $d$  stocks. On day  $t$ , stock  $i$  has gross return  $r_t(i) = \text{price}_{t+1}(i)/\text{price}_t(i) \geq 0$ .

A portfolio is  $p_t \in \Delta_d$  (wealth fractions), rebalanced daily. Starting from wealth  $v_1$ ,

$$v_{T+1} = v_1 \prod_{t=1}^T \langle p_t, r_t \rangle,$$

so the log-wealth is  $\log(v_{T+1}/v_1) = \sum_{t=1}^T \log \langle p_t, r_t \rangle$ .

**Classical (Markowitz):** assume  $r_t$  IID with mean  $\mu$  and covariance  $\Sigma$  estimated from history; solve  $\min_{p \in \Delta_d} p^\top \Sigma p$  s.t.  $p^\top \mu \geq \bar{r}$ . IID fails exactly when it matters.

## Online view

Treat  $r_1, r_2, \dots$  as adversarial. Loss on day  $t$ :  $\ell_t(p) = -\log \langle p, r_t \rangle$ . Compete with a fixed comparator:

$$\max_{p \in \mathcal{C}} \sum_{t=1}^T \log \langle p, r_t \rangle - \sum_{t=1}^T \log \langle p_t, r_t \rangle.$$

Which comparator class  $\mathcal{C}$ ?

- $\mathcal{C} = \{e_1, \dots, e_d\}$ : the best single stock
- $\mathcal{C} = \Delta_d$ : the best fixed (constantly rebalanced) portfolio

# Competing with the best single stock

## Theorem

Run exponential weights over the  $d$  stocks with  $\eta = 1$  and losses  $\ell_t(i) = -\log r_t(i)$ . Then for any return sequence,

$$\max_{i \in [d]} \sum_{t=1}^T \log r_t(i) - \sum_{t=1}^T \log \langle p_t, r_t \rangle \leq \log d.$$

# Competing with the best single stock

## Theorem

Run exponential weights over the  $d$  stocks with  $\eta = 1$  and losses  $\ell_t(i) = -\log r_t(i)$ . Then for any return sequence,

$$\max_{i \in [d]} \sum_{t=1}^T \log r_t(i) - \sum_{t=1}^T \log \langle p_t, r_t \rangle \leq \log d.$$

**What the algorithm is doing here.** With  $\eta = 1$  the update  $w_{t+1}(i) = w_t(i) e^{-\ell_t(i)}$  reads

$$w_{t+1}(i) = w_t(i) r_t(i), \quad \text{so} \quad w_{t+1}(i) = \prod_{s=1}^t r_s(i) \quad (\text{starting from } w_1(i) = 1),$$

the wealth of one dollar put into stock  $i$  on day 1. The portfolio is  $p_t(i) = w_t(i)/W_t$  with  $W_t = \sum_{i=1}^d w_t(i)$ : each stock gets a share proportional to how much it has grown so far.

# Competing with the best single stock

## Theorem

Run exponential weights over the  $d$  stocks with  $\eta = 1$  and losses  $\ell_t(i) = -\log r_t(i)$ . Then for any return sequence,

$$\max_{i \in [d]} \sum_{t=1}^T \log r_t(i) - \sum_{t=1}^T \log \langle p_t, r_t \rangle \leq \log d.$$

**What the algorithm is doing here.** With  $\eta = 1$  the update  $w_{t+1}(i) = w_t(i) e^{-\ell_t(i)}$  reads

$$w_{t+1}(i) = w_t(i) r_t(i), \quad \text{so} \quad w_{t+1}(i) = \prod_{s=1}^t r_s(i) \quad (\text{starting from } w_1(i) = 1),$$

the wealth of one dollar put into stock  $i$  on day 1. The portfolio is  $p_t(i) = w_t(i)/W_t$  with  $W_t = \sum_{i=1}^d w_t(i)$ : each stock gets a share proportional to how much it has grown so far.

**Why the usual theorem does not apply directly.** The losses  $-\log r_t(i)$  are not in  $[0, 1]$  (they are unbounded in both directions), so the bound  $\log d/\eta + \eta T/8$  is not available. We redo the potential argument from scratch, and it comes out **cleaner**: no  $\eta T/8$  term at all.

## Competing with the best single stock: proof

**Step 1: the potential ratio is one day's wealth factor.** Using  $w_{t+1}(i) = w_t(i) r_t(i)$  and  $p_t(i) = w_t(i)/W_t$ ,

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

# Competing with the best single stock: proof

**Step 1: the potential ratio is one day's wealth factor.** Using  $w_{t+1}(i) = w_t(i) r_t(i)$  and  $p_t(i) = w_t(i)/W_t$ ,

$$\frac{W_{t+1}}{W_t} = \frac{1}{W_t} \sum_{i=1}^d w_{t+1}(i) = \sum_{i=1}^d \frac{w_t(i)}{W_t} r_t(i) = \langle p_t, r_t \rangle.$$

**Step 2: telescope.** Multiplying over the  $T$  days, with  $W_1 = \sum_i w_1(i) = d$ ,

$$\prod_{t=1}^T \langle p_t, r_t \rangle = \prod_{t=1}^T \frac{W_{t+1}}{W_t} = \frac{W_{T+1}}{W_1} = \frac{1}{d} \sum_{i=1}^d \prod_{t=1}^T r_t(i).$$

**Step 3: an average is at least  $1/d$  of its largest term.** For  $i^*$  the best stock in hindsight,  $\sum_i \prod_t r_t(i) \geq \prod_t r_t(i^*)$ , so

$$\sum_{t=1}^T \log \langle p_t, r_t \rangle = \log \frac{W_{T+1}}{d} \geq \log \frac{\prod_t r_t(i^*)}{d} = \sum_{t=1}^T \log r_t(i^*) - \log d.$$

# Competing with the best single stock: proof

**Step 1: the potential ratio is one day's wealth factor.** Using  $w_{t+1}(i) = w_t(i) r_t(i)$  and  $p_t(i) = w_t(i)/W_t$ ,

$$\frac{W_{t+1}}{W_t} = \frac{1}{W_t} \sum_{i=1}^d w_{t+1}(i) = \sum_{i=1}^d \frac{w_t(i)}{W_t} r_t(i) = \langle p_t, r_t \rangle.$$

**Step 2: telescope.** Multiplying over the  $T$  days, with  $W_1 = \sum_i w_1(i) = d$ ,

$$\prod_{t=1}^T \langle p_t, r_t \rangle = \prod_{t=1}^T \frac{W_{t+1}}{W_t} = \frac{W_{T+1}}{W_1} = \frac{1}{d} \sum_{i=1}^d \prod_{t=1}^T r_t(i).$$

**Step 3: an average is at least  $1/d$  of its largest term.** For  $i^*$  the best stock in hindsight,  $\sum_i \prod_t r_t(i) \geq \prod_t r_t(i^*)$ , so

$$\sum_{t=1}^T \log \langle p_t, r_t \rangle = \log \frac{W_{T+1}}{d} \geq \log \frac{\prod_t r_t(i^*)}{d} = \sum_{t=1}^T \log r_t(i^*) - \log d. \quad \blacksquare$$

**Versus the exponential weights proof:** same potential, same sandwich, but Step 1 there needed Hoeffding's lemma ( $\eta^2/8$  slack per round); here  $e^{-\ell_t(i)} = r_t(i)$  is linear, so  $\mathbb{E}_{I \sim p_t}[e^{-\ell_t(I)}] = \langle p_t, r_t \rangle$  exactly.

# The loss changes the comparator

**Linear losses**  $p \mapsto \langle p, \ell_t \rangle$  (classification, experts):

$$\min_{p \in \Delta_d} \sum_t \langle p, \ell_t \rangle = \min_{i \in [d]} \sum_t \ell_t(i).$$

A linear function on the simplex is minimized at a vertex. Competing with the best mixture is the same as competing with the best expert.

Nothing gained.

**Log losses**  $p \mapsto -\log \langle p, r_t \rangle$  are convex but not linear.

# The loss changes the comparator

**Linear losses**  $p \mapsto \langle p, \ell_t \rangle$  (classification, experts):

$$\min_{p \in \Delta_d} \sum_t \langle p, \ell_t \rangle = \min_{i \in [d]} \sum_t \ell_t(i).$$

A linear function on the simplex is minimized at a vertex. Competing with the best mixture is the same as competing with the best expert.

Nothing gained.

**Log losses**  $p \mapsto -\log \langle p, r_t \rangle$  are convex but not linear.

## Example: $d = 2$ , alternating returns

$r_t = (2, \frac{1}{2})$  for odd  $t$ ,  $r_t = (\frac{1}{2}, 2)$  for even  $t$ .

- Each single stock:  $\times 2$  then  $\times \frac{1}{2}$ . Wealth after  $T$  days: 1.
- Uniform  $p = (\frac{1}{2}, \frac{1}{2})$ :  $\langle p, r_t \rangle = \frac{5}{4}$  every day. Wealth:  $(5/4)^T$ .

The best fixed *portfolio* beats the best stock by an exponential factor.

# The loss changes the comparator

**Linear losses**  $p \mapsto \langle p, \ell_t \rangle$  (classification, experts):

$$\min_{p \in \Delta_d} \sum_t \langle p, \ell_t \rangle = \min_{i \in [d]} \sum_t \ell_t(i).$$

A linear function on the simplex is minimized at a vertex. Competing with the best mixture is the same as competing with the best expert.

Nothing gained.

**Log losses**  $p \mapsto -\log \langle p, r_t \rangle$  are convex but not linear.

**Example:  $d = 2$ , alternating returns**

$r_t = (2, \frac{1}{2})$  for odd  $t$ ,  $r_t = (\frac{1}{2}, 2)$  for even  $t$ .

- Each single stock:  $\times 2$  then  $\times \frac{1}{2}$ . Wealth after  $T$  days: 1.
- Uniform  $p = (\frac{1}{2}, \frac{1}{2})$ :  $\langle p, r_t \rangle = \frac{5}{4}$  every day. Wealth:  $(5/4)^T$ .

The best fixed *portfolio* beats the best stock by an exponential factor.

To compete with the best  $p \in \Delta_d$  we need exponential weights over a *continuum* of hypotheses.

# Exponential weights over a continuum

## Continuous exponential weights (CEW)

**Input:** convex  $\mathcal{A} \subseteq \mathbb{R}^d$ ,  $\eta > 0$ , base measure  $\nu$  on  $\mathcal{A}$   
**for**  $t = 1, 2, \dots, T$ :

Sample  $a_t \sim p_t$ ,  $dp_t(a) \propto e^{-\eta \sum_{s < t} \ell_s(a)} d\nu(a)$

Observe  $\ell_t$ ; suffer  $\ell_t(a_t)$

## Theorem (CEW regret)

If each  $\ell_t : \mathcal{A} \rightarrow [0, 1]$  is convex and  $\nu$  is Lebesgue, the expected regret  $\mathbb{E}[\sum_t \ell_t(a_t)] - \min_{a \in \mathcal{A}} \sum_t \ell_t(a)$  is at most

$$\frac{d \log T}{\eta} + \frac{\eta T}{8} + 1 \stackrel{\eta = \sqrt{8d \log T / T}}{=} \sqrt{\frac{dT \log T}{2}} + 1.$$

# Exponential weights over a continuum

## Continuous exponential weights (CEW)

**Input:** convex  $\mathcal{A} \subseteq \mathbb{R}^d$ ,  $\eta > 0$ , base measure  $\nu$  on  $\mathcal{A}$   
**for**  $t = 1, 2, \dots, T$ :

Sample  $a_t \sim p_t$ ,  $dp_t(a) \propto e^{-\eta \sum_{s < t} \ell_s(a)} d\nu(a)$

Observe  $\ell_t$ ; suffer  $\ell_t(a_t)$

## Theorem (CEW regret)

If each  $\ell_t : \mathcal{A} \rightarrow [0, 1]$  is convex and  $\nu$  is Lebesgue, the expected regret  $\mathbb{E}[\sum_t \ell_t(a_t)] - \min_{a \in \mathcal{A}} \sum_t \ell_t(a)$  is at most

$$\frac{d \log T}{\eta} + \frac{\eta T}{8} + 1 \stackrel{\eta = \sqrt{8d \log T / T}}{=} \sqrt{\frac{dT \log T}{2}} + 1.$$

**What changes in the proof.** The upper bound on  $\log(W_{T+1}/W_1)$  is identical. For the lower bound a single point  $a^*$  has measure zero, so integrate over the shrunken copy  $\mathcal{A}^* = (1 - \frac{1}{T})a^* + \frac{1}{T}\mathcal{A}$ : by convexity  $\ell_t(a) \leq \ell_t(a^*) + \frac{1}{T}$  on  $\mathcal{A}^*$ , and  $\nu(\mathcal{A}^*)/\nu(\mathcal{A}) = T^{-d}$ . Hence

$$\log \frac{W_{T+1}}{W_1} \geq \log \frac{\nu(\mathcal{A}^*)}{\nu(\mathcal{A})} - \eta \sum_{t=1}^T (\ell_t(a^*) + \frac{1}{T}) = -d \log T - \eta \sum_{t=1}^T \ell_t(a^*) - \eta,$$

and combining with the upper bound as before gives the theorem:  $\log(\text{number of actions})$  has become  $d \log T$ , the log-number of hypotheses at resolution  $1/T$  in  $d$  dimensions. ■

# Universal portfolio (Cover, 1991)

Take  $\mathcal{A} = \Delta_d$ ,  $\ell_t(p) = -\log\langle p, r_t \rangle$ ,  $\eta = 1$ :  $dp_t(p) \propto \prod_{s < t} \langle p, r_s \rangle dp$ , the wealth the rebalanced portfolio  $p$  would have earned so far.

## Cover's algorithm

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

Play the posterior mean  $\bar{p}_t = \frac{\int_{\Delta_d} p \prod_{s < t} \langle p, r_s \rangle dp}{\int_{\Delta_d} \prod_{s < t} \langle p, r_s \rangle dp}$

Observe  $r_t$ ; wealth multiplies by  $\langle \bar{p}_t, r_t \rangle$

No sampling:  $\mathbb{E}_{p \sim p_t} \langle p, r_t \rangle = \langle \bar{p}_t, r_t \rangle$  since the wealth factor is *linear* in  $p$ .

## Corollary

For any return sequence,

$$\max_{p \in \Delta_d} \sum_t \log \langle p, r_t \rangle - \sum_t \log \langle \bar{p}_t, r_t \rangle \leq d \log T + 2.$$

# Universal portfolio (Cover, 1991)

Take  $\mathcal{A} = \Delta_d$ ,  $\ell_t(p) = -\log\langle p, r_t \rangle$ ,  $\eta = 1$ :  $dp_t(p) \propto \prod_{s < t} \langle p, r_s \rangle dp$ , the wealth the rebalanced portfolio  $p$  would have earned so far.

## Cover's algorithm

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

Play the posterior mean  $\bar{p}_t = \frac{\int_{\Delta_d} p \prod_{s < t} \langle p, r_s \rangle dp}{\int_{\Delta_d} \prod_{s < t} \langle p, r_s \rangle dp}$

Observe  $r_t$ ; wealth multiplies by  $\langle \bar{p}_t, r_t \rangle$

## Corollary

For any return sequence,

$$\max_{p \in \Delta_d} \sum_t \log \langle p, r_t \rangle - \sum_t \log \langle \bar{p}_t, r_t \rangle \leq d \log T + 2.$$

No sampling:  $\mathbb{E}_{p \sim p_t} \langle p, r_t \rangle = \langle \bar{p}_t, r_t \rangle$  since the wealth factor is *linear* in  $p$ .

**Proof.**  $\langle \bar{p}_t, r_t \rangle = W_{t+1}/W_t$  exactly (linearity), so  $\sum_t \log \langle \bar{p}_t, r_t \rangle = \log(W_{T+1}/W_1)$ . On  $\mathcal{A}^* = (1 - \frac{1}{T})p^* + \frac{1}{T}\Delta_d$  we have  $\langle p, r_t \rangle \geq (1 - \frac{1}{T})\langle p^*, r_t \rangle$ , so

$$\log \frac{W_{T+1}}{W_1} \geq -d \log T + \sum_{t=1}^T \log \langle p^*, r_t \rangle + T \log \left(1 - \frac{1}{T}\right) \geq \sum_{t=1}^T \log \langle p^*, r_t \rangle - d \log T - 2. \quad \blacksquare$$

**Cost:** an integral over  $\Delta_d$  every round; exact only for tiny  $d$ , otherwise sampling.

# 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.  $\hat{L}_n(\hat{f}) - \hat{L}_n(f^*) \leq 0$ .

### 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  $\leq 0$ , 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  $\leq 0$ , 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  $\leq 0$ , 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  $\epsilon$  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.