From stochastic mixability to fast rates

Nishant Mehta

THE AUSTRALIAN NATIONAL UNIVERSITY

Bob Williamson

THE AUSTRALIAN NATIONAL UNIVERSITY

NICTA

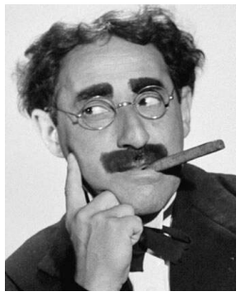
Prediction with expert advice

In round $t = 1, 2, \ldots, T$

Prediction with expert advice

In round $t = 1, 2, \ldots, T$

1. K experts predict $a_t^{(1)}, \ldots, a_t^{(K)}$



AMZN

GOOG

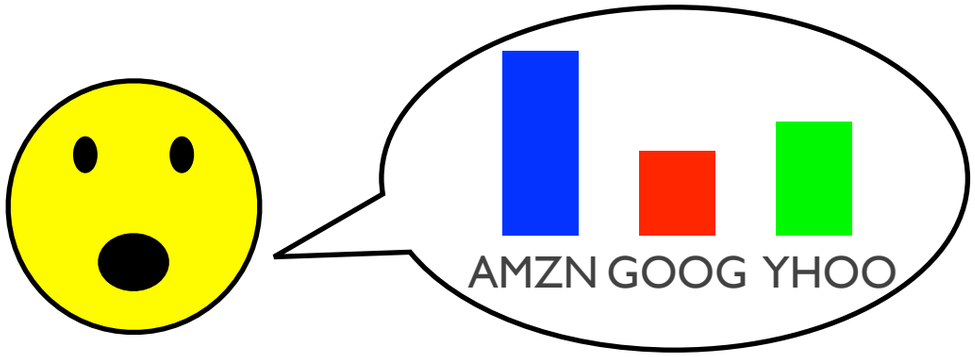
YHOO

Prediction with expert advice

In round $t = 1, 2, \ldots, T$

I K experts predict $a_t^{(1)}, \ldots, a_t^{(K)}$

II Learner predicts a_t

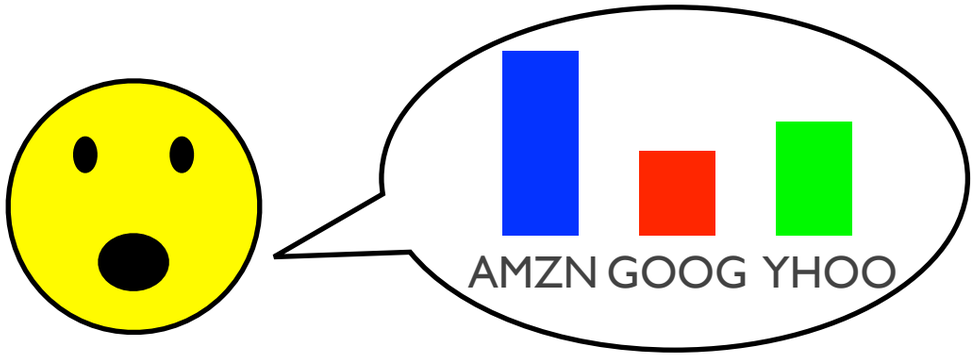


Prediction with expert advice

In round $t = 1, 2, \ldots, T$

I K experts predict $a_t^{(1)}, \ldots, a_t^{(K)}$

II Learner predicts a_t



III Nature reveals outcome ω_t

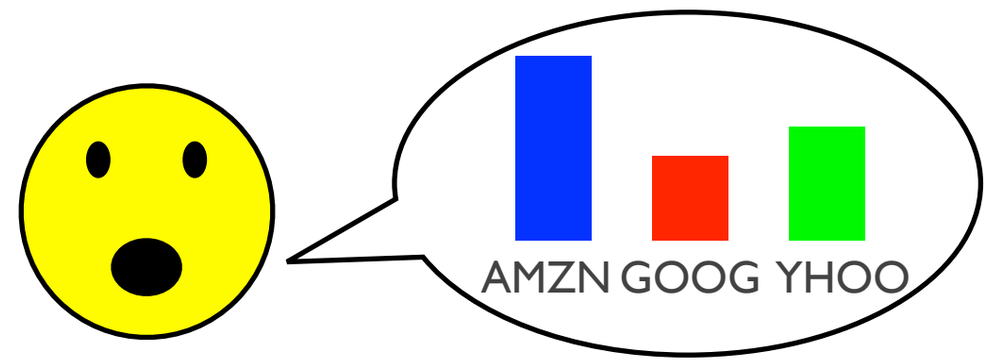
AMZN	GOOG	YHOO
-3%	+2%	+5%

Prediction with expert advice

In round $t = 1, 2, \ldots, T$

I K experts predict $a_t^{(1)}, \ldots, a_t^{(K)}$

II Learner predicts a_t



III Nature reveals outcome ω_t

AMZN	GOOG	YHOO
-3%	+2%	+5%

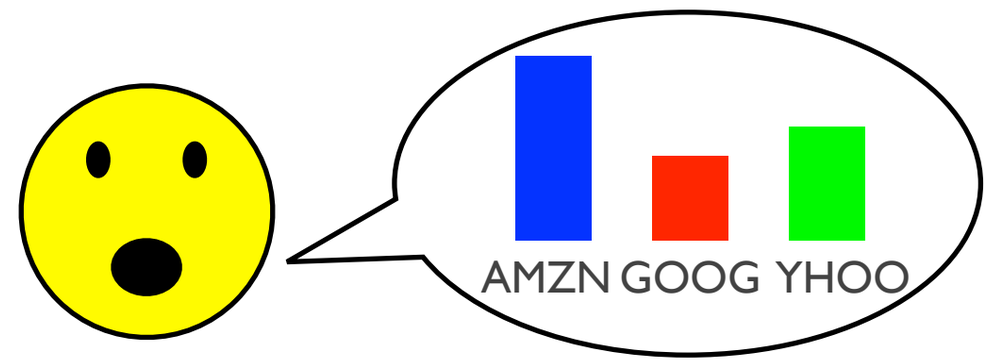
IV Learner/experts suffer loss $\ell(\omega_t, a)$

Prediction with expert advice

In round $t = 1, 2, \ldots, T$

I K experts predict $a_t^{(1)}, \ldots, a_t^{(K)}$

II Learner predicts a_t



III Nature reveals outcome ω_t

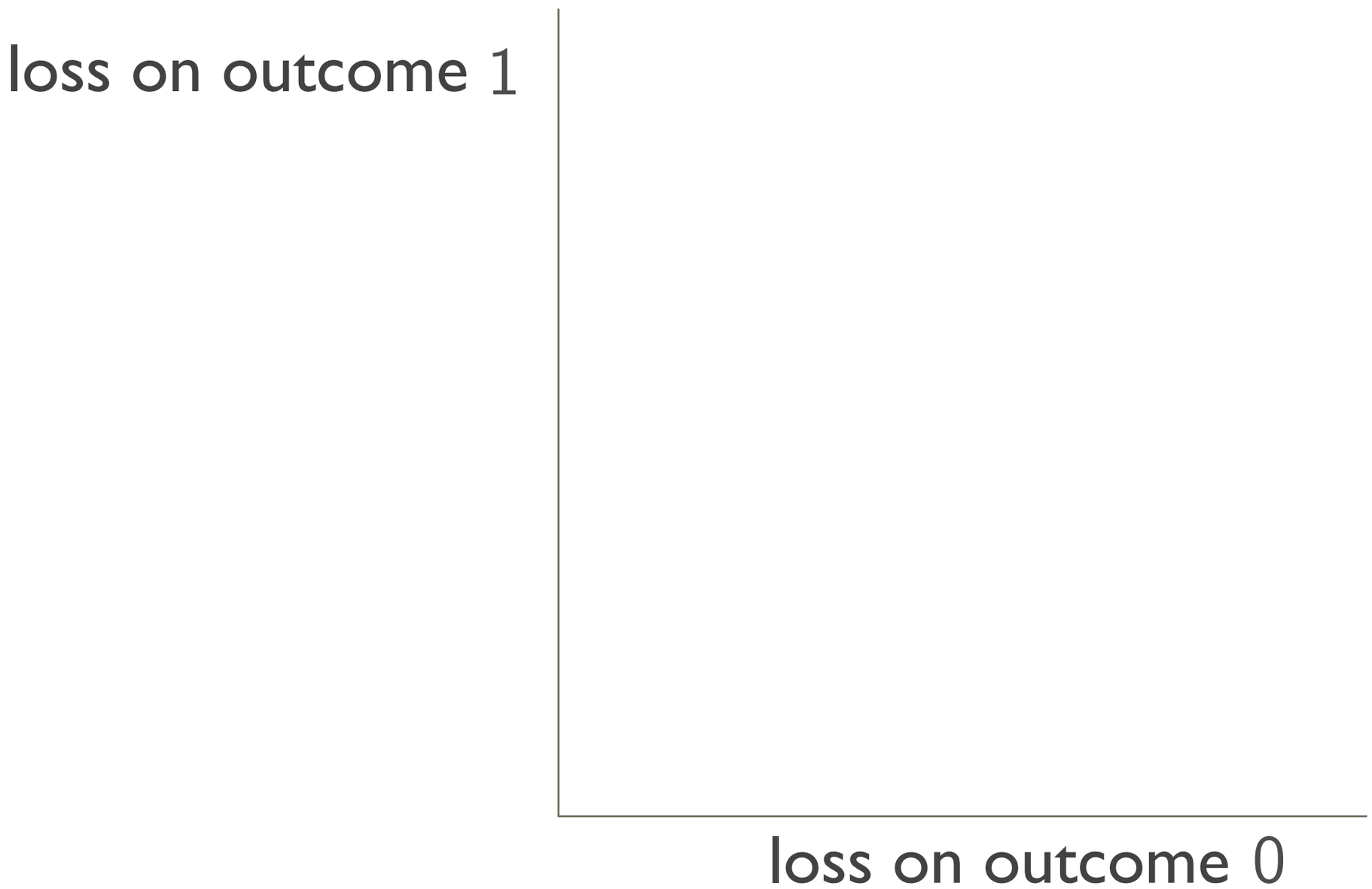
AMZN	GOOG	YHOO
-3%	+2%	+5%

IV Learner/experts suffer loss $\ell(\omega_t, a)$

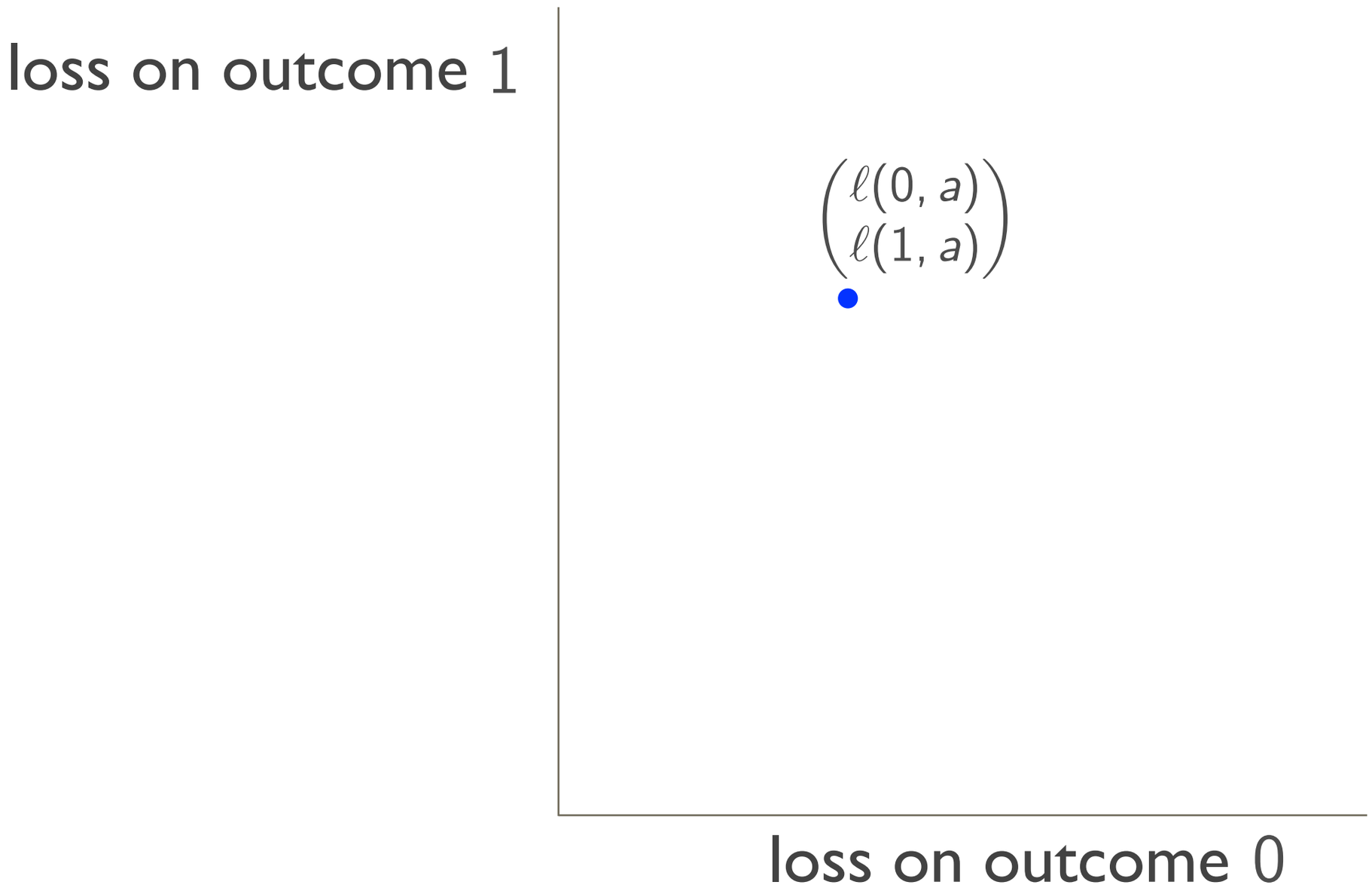
GOAL: MINIMIZE THE REGRET

$$\sum_{t=1}^{T} \ell(\omega_t, a_t) - \min_{j \in [k]} \sum_{t=1}^{T} \ell(\omega_t, a_t^{(j)})$$

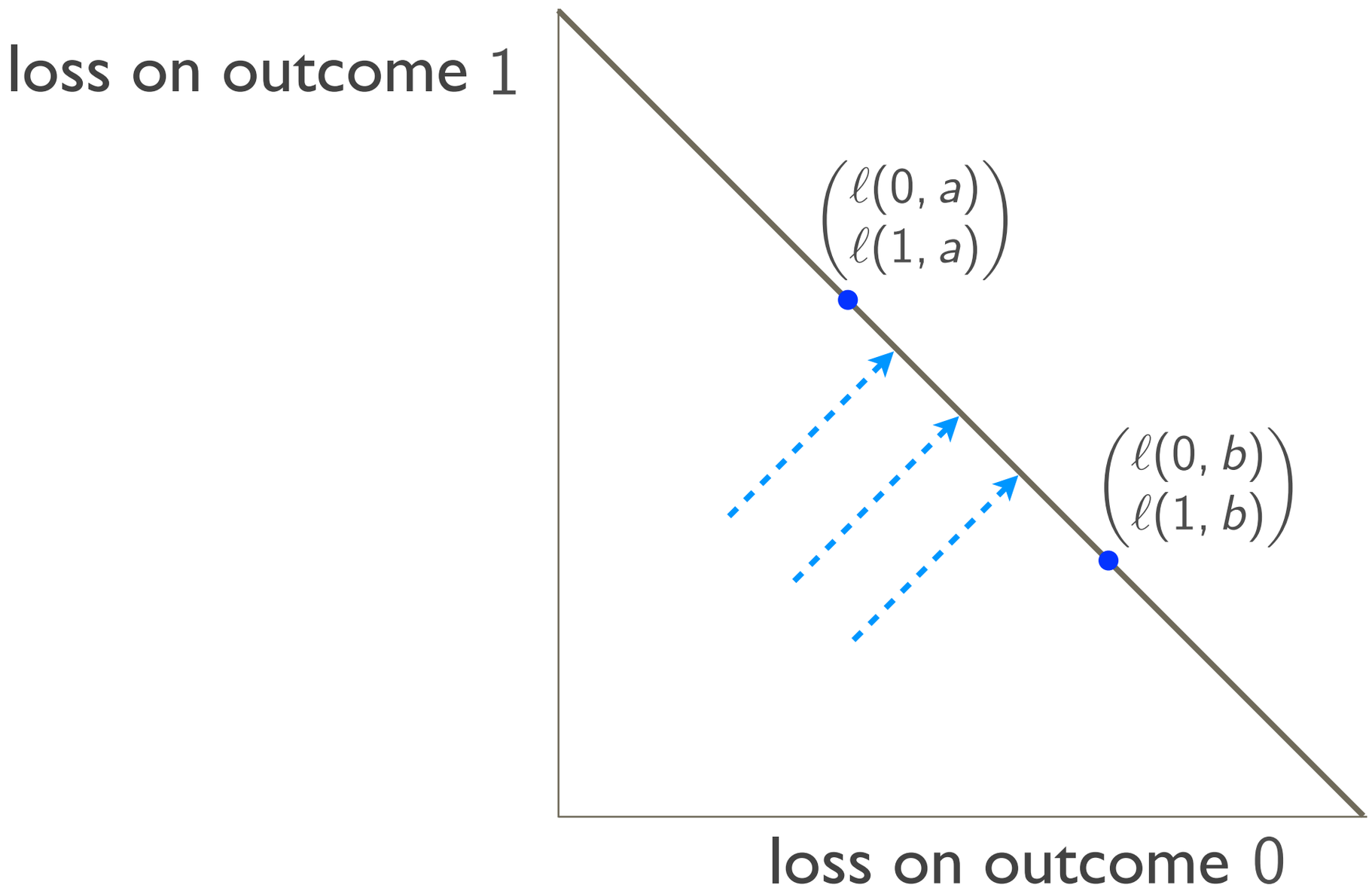
Absolute loss lower bound



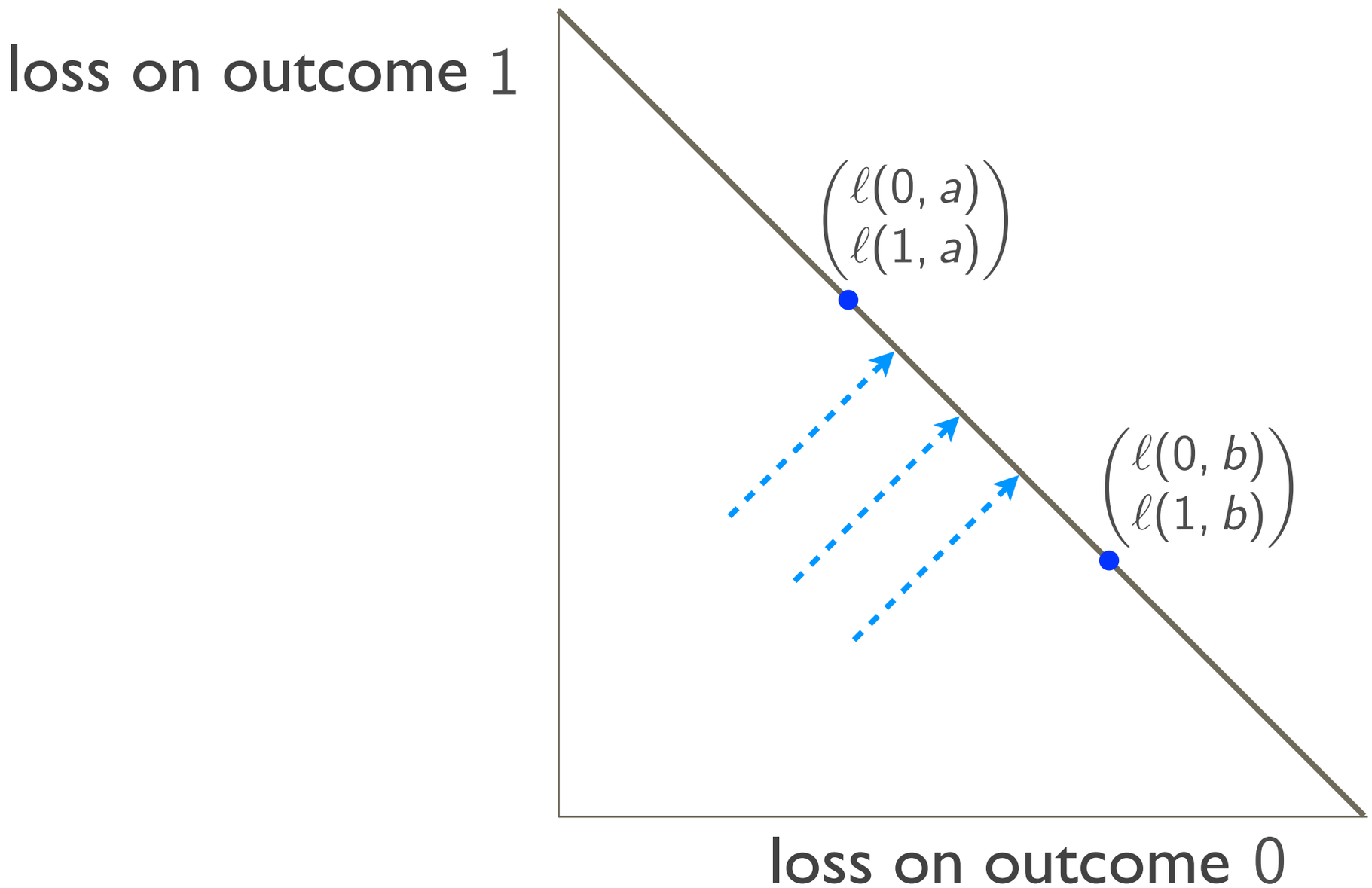
Absolute loss lower bound



Absolute loss lower bound



Absolute loss lower bound

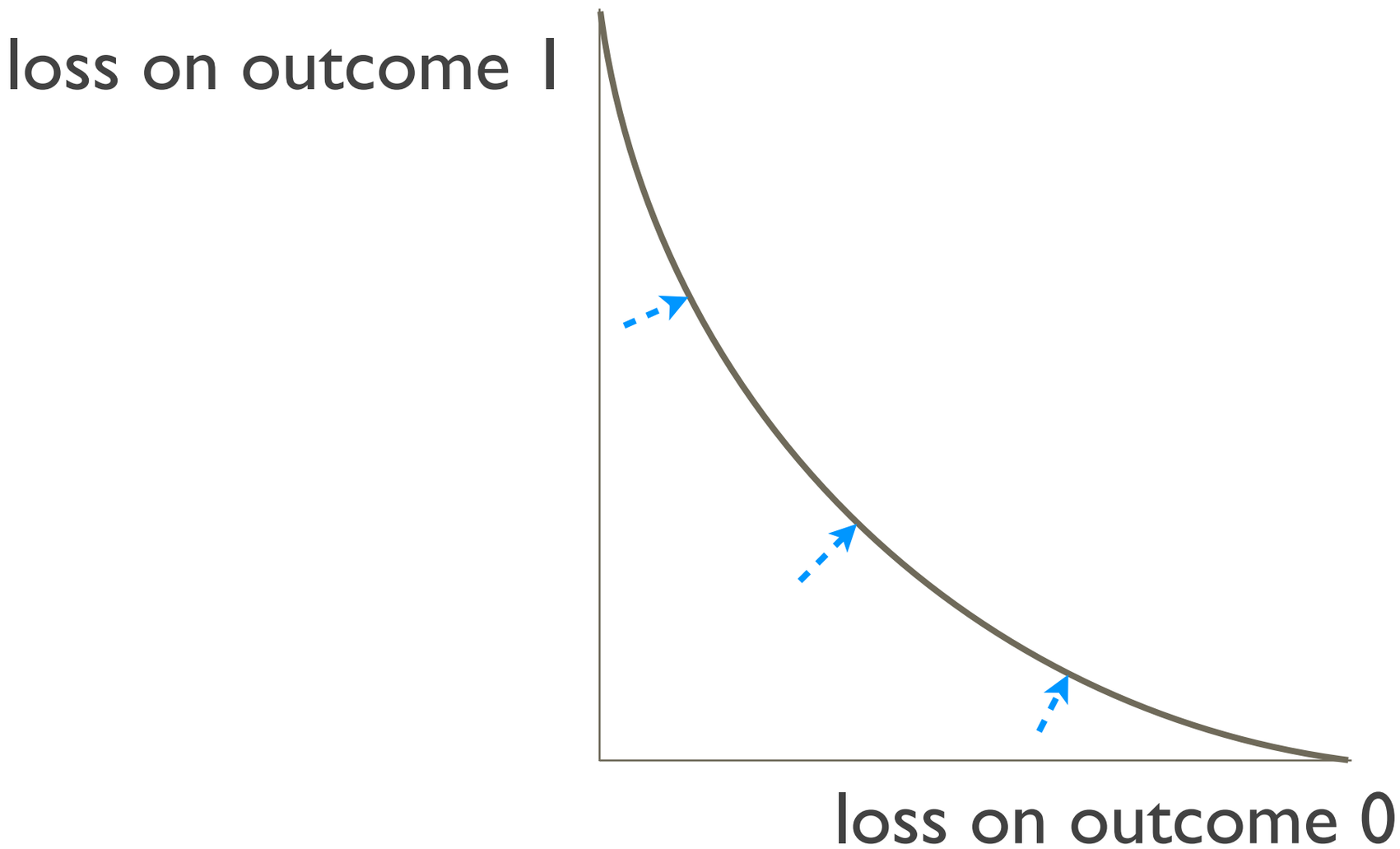


FLATNESS IS BAD!

Absolute loss: Minimax regret is $\Theta(\sqrt{T \log K})$

[Cesa-Bianchi et al., 1997]

Mixable losses break lower bound



η-mixable losses: Minimax regret is $\frac{\log K}{\eta}$ constant in T!

[Vovk, 1990]

η-Mixability

For each distribution over actions π

There is an action a^* such that

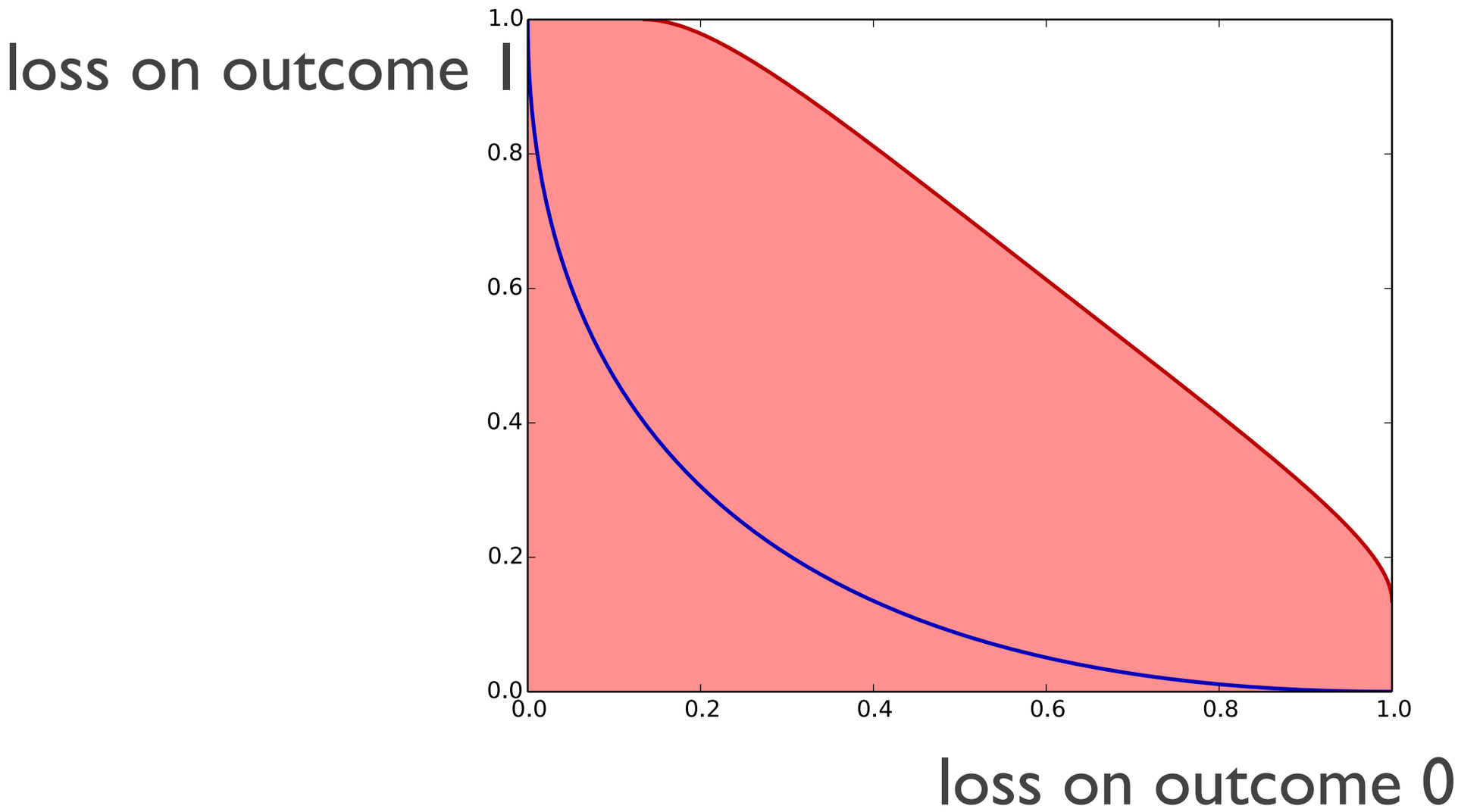
For all outcomes ω: $\ell(\omega, a^*) \leq \underbrace{-\frac{1}{\eta} \log \mathrm{E}_{a \sim \pi}\, e^{-\eta \ell(\omega, a)}}_{\ell_{\eta\text{-mix}}(\omega, \pi)}$

η-mixable losses: Minimax regret is $\frac{\log K}{\eta}$ constant in T!

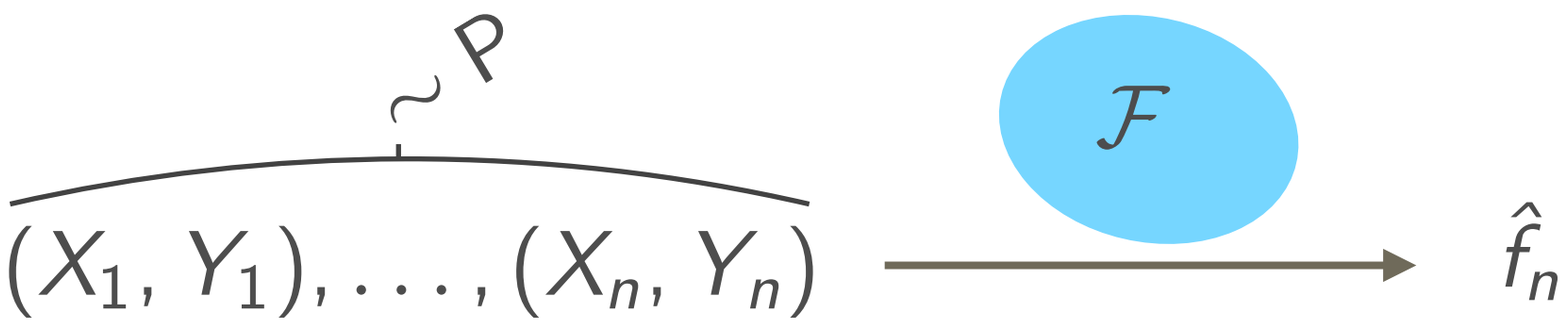
[Vovk, 1998]

Squared loss mixability

$$\begin{pmatrix} \ell_0 \\ \ell_1 \end{pmatrix} \mapsto \begin{pmatrix} e^{-\eta \ell_0} \\ e^{-\eta \ell_1} \end{pmatrix}$$



Regret in statistical learning



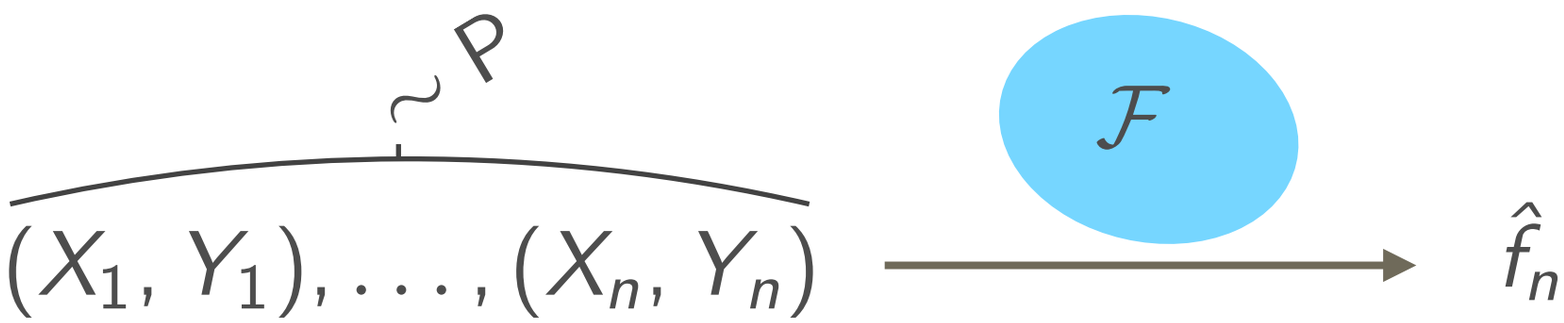
Regret in statistical learning

$$\overbrace{(X_1, Y_1), \ldots, (X_n, Y_n)}^{\sim \mathrm{P}} \xrightarrow{\mathcal{F}} \hat{f}_n$$

estimator

$$\mathrm{E}_{(X,Y)\sim \mathrm{P}}\, \ell\big((X, Y), \hat{f}_n\big) - \inf_{f \in \mathcal{F}} \mathrm{E}_{(X,Y)\sim \mathrm{P}}\, \ell\big((X, Y), f\big) = O\left(\frac{1}{\sqrt{n}}\right)$$

Regret in statistical learning



estimator

$$\mathsf{E}_{(X,Y)\sim\mathsf{P}}\,\ell\big((X,Y),\hat{f}_n\big) - \inf_{f\in\mathcal{F}} \mathsf{E}_{(X,Y)\sim\mathsf{P}}\,\ell\big((X,Y),f\big) = O\big(\tfrac{1}{\sqrt{n}}\big)$$

THIS TALK: Under *Stochastic Mixability*

$$\mathsf{E}_{(X,Y)\sim\mathsf{P}}\,\ell\big((X,Y),\hat{f}_n\big) - \inf_{f\in\mathcal{F}} \mathsf{E}_{(X,Y)\sim\mathsf{P}}\,\ell\big((X,Y),f\big) = O\big(\tfrac{1}{n}\big)$$

Online Learning $\longrightarrow$ Statistical Learning

	Prediction with Expert Advice (constant experts)	Statistical Learning
Actions	$a \in \mathcal{A}$	$f \in \mathcal{F}$
Outcomes	$\omega \in \Omega$	$(X, Y) \in \mathcal{X} \times \mathcal{Y}$
Loss	$\ell(\omega, a)$	$\ell((X, Y), f)$
Nature's Strategy	$P(\omega_1, \ldots, \omega_T)$ (full joint)	$P\big((X_j, Y_j)_{j=1}^n\big) = P^n\big((X, Y)\big)$ (i.i.d.)

Online Learning $\longrightarrow$ Statistical Learning

	Prediction with Expert Advice (constant experts)	Statistical Learning
Actions	$a \in \mathcal{A}$ $(\mathcal{A} = [0, 1])$	$f \in \mathcal{F}$ $(\mathcal{F} = \text{linear separators})$
Outcomes	$\omega \in \Omega$ $(\Omega = \{0, 1\})$	$(X, Y) \in \mathcal{X} \times \mathcal{Y}$ $(\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \{0, 1\})$
Loss	$\ell(\omega, a)$ $(\lvert\omega - a\rvert)$	$\ell((X, Y), f)$ $(\ell(Y, f(X)) = \lvert Y - f(X)\rvert)$
Nature's Strategy	$P(\omega_1, \ldots, \omega_T)$ (full joint)	$P\big((X_j, Y_j)_{j=1}^n\big) = P^n\big((X, Y)\big)$ (i.i.d.)

η-Mixability $\longrightarrow$ η-Stochastic Mixability

DEFINITION [Van Erven et al.]

A problem $(\ell, \mathcal{F}, \mathrm{P})$ is η-stochastically mixable if

there is some $f^* \in \mathcal{F}$ such that for all distributions π over $\mathcal{F}$:

$$\mathrm{E}_{(X,Y)\sim \mathrm{P}}\, \ell\big((X, Y), f^*\big) \leq \underbrace{\mathrm{E}_{(X,Y)\sim \mathrm{P}} -\frac{1}{\eta} \log \mathrm{E}_{f\sim\pi}\, e^{-\eta \ell\big((X,Y),f\big)}}_{\mathrm{E}_{(X,Y)\sim \mathrm{P}}\, \ell_{\eta\text{-mix}}\big((X, Y), \pi\big)}$$

CGF view of Stochastic Mixability

For all distributions π over $\mathcal{F}$:

$$\mathrm{E}\,\ell\big((X,Y),f^*\big) \le \mathrm{E}\,\ell_{\eta\text{-mix}}\big((X,Y),\pi\big)$$

EQUIVALENT! [Van Erven et al.]

For all $f \in \mathcal{F}$:

$$\log \mathrm{E}\exp\Big(-\eta\big[\ell\big(Y,f(X)\big) - \ell\big(Y,f^*(X)\big)\big]\Big) \le 0$$

CGF view of Stochastic Mixability

For all distributions π over $\mathcal{F}$:

$$\mathrm{E}\,\ell\big((X,Y), f^*\big) \leq \mathrm{E}\,\ell_{\eta\text{-mix}}\big((X,Y), \pi\big)$$

EQUIVALENT! [Van Erven et al.]

For all $f \in \mathcal{F}$:

$$\log \mathrm{E} \exp\Big(-\eta\big[\overbrace{\ell\big(Y, f(X)\big) - \ell\big(Y, f^*(X)\big)}^{\text{excess loss } Z_f}\big]\Big) \leq 0$$

CGF view of Stochastic Mixability

For all distributions π over $\mathcal{F}$:

$$\mathrm{E}\,\ell\big((X,Y),f^*\big) \le \mathrm{E}\,\ell_{\eta\text{-mix}}\big((X,Y),\pi\big)$$

EQUIVALENT! [Van Erven et al.]

For all $f \in \mathcal{F}$:

$$\underbrace{\log \mathrm{E} \exp\Big(-\eta\big[\overbrace{\ell\big(Y,f(X)\big)-\ell\big(Y,f^*(X)\big)}^{\text{excess loss } Z_f}\big]\Big)}_{\Lambda_{-Z_f}(\eta)} \le 0$$

Fast rates from stochastic mixability

THIS TALK: LOSSES IN [0, 1] FOR SIMPLICITY

With probability at least $1 - \delta$, under stochastic mixability

Finite classes

$$R(\hat{f}_{\mathrm{ERM}}) - R(f^*) \leq \frac{6\left(\log \frac{1}{\delta} + \log |\mathcal{F}|\right)}{n\eta^*}$$

Fast rates from stochastic mixability

THIS TALK: LOSSES IN [0, 1] FOR SIMPLICITY

With probability at least $1 - \delta$, under stochastic mixability

Finite classes

$$R(\hat{f}_{\mathrm{ERM}}) - R(f^*) \leq \frac{6\left(\log \frac{1}{\delta} + \log |\mathcal{F}|\right)}{n\eta^*}$$

VC-type classes

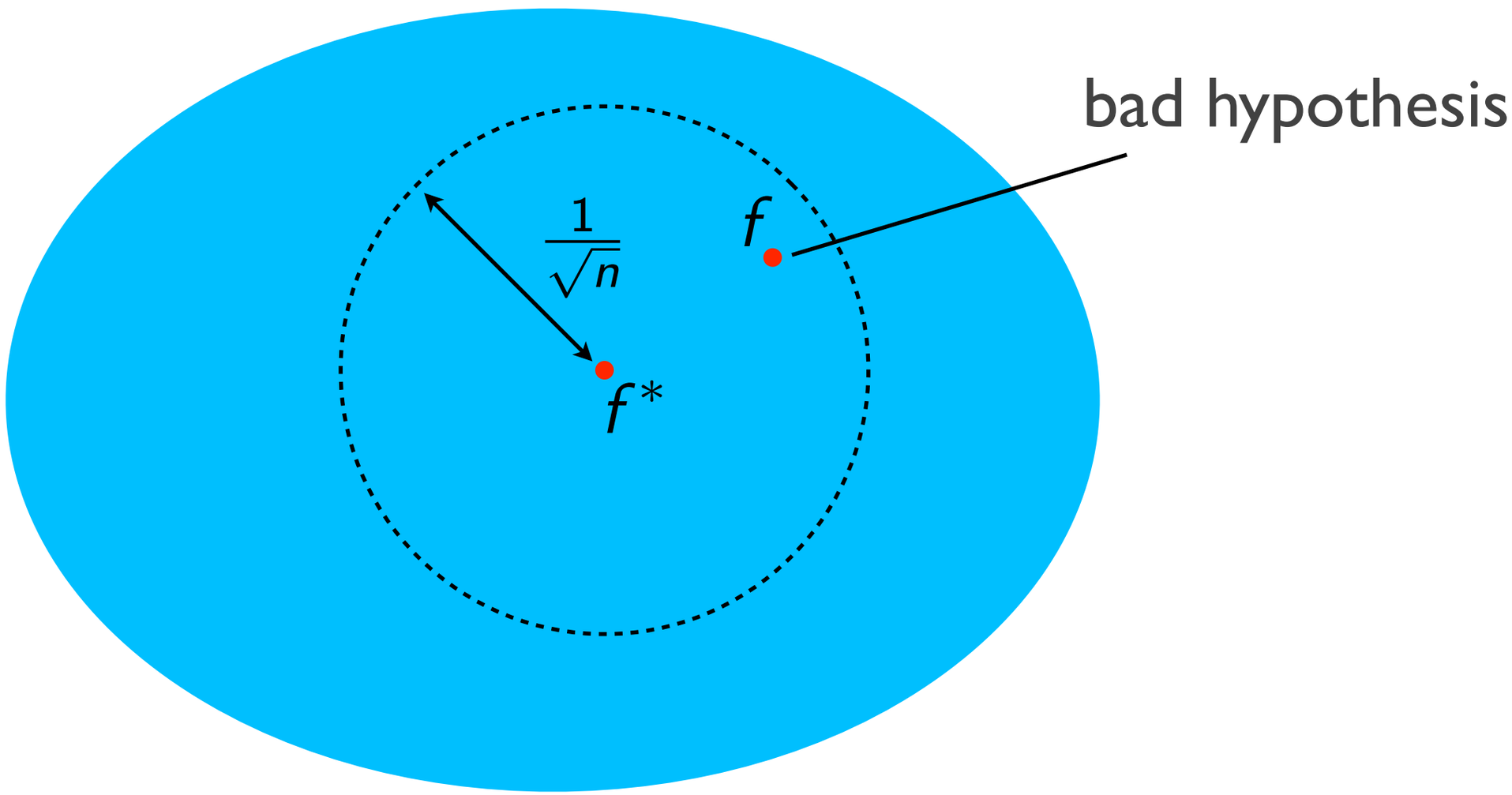
$$R(\hat{f}_{\mathrm{ERM}}) - R(f^*) = O\left(\frac{\left(\sqrt{\mathrm{VC}(\mathcal{F}) \log n} + \sqrt{\log \frac{1}{\delta}}\right)^2}{n\eta^*}\right)$$

Avoiding Bad Hypotheses

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f(X_j)) \leq \frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f^*(X_j))\right)$$

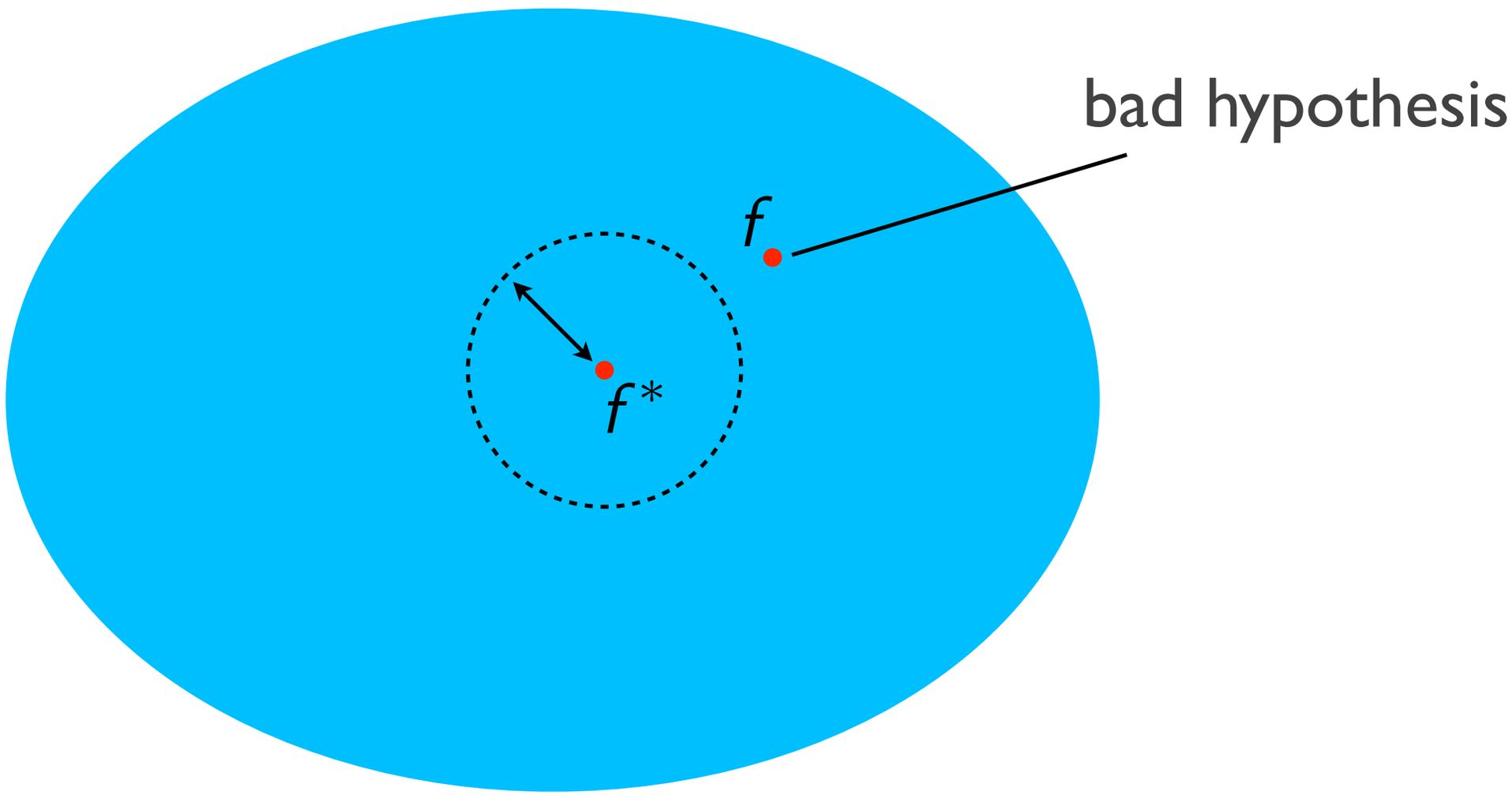
Avoiding Bad Hypotheses

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f(X_j)) \leq \frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f^*(X_j))\right)$$



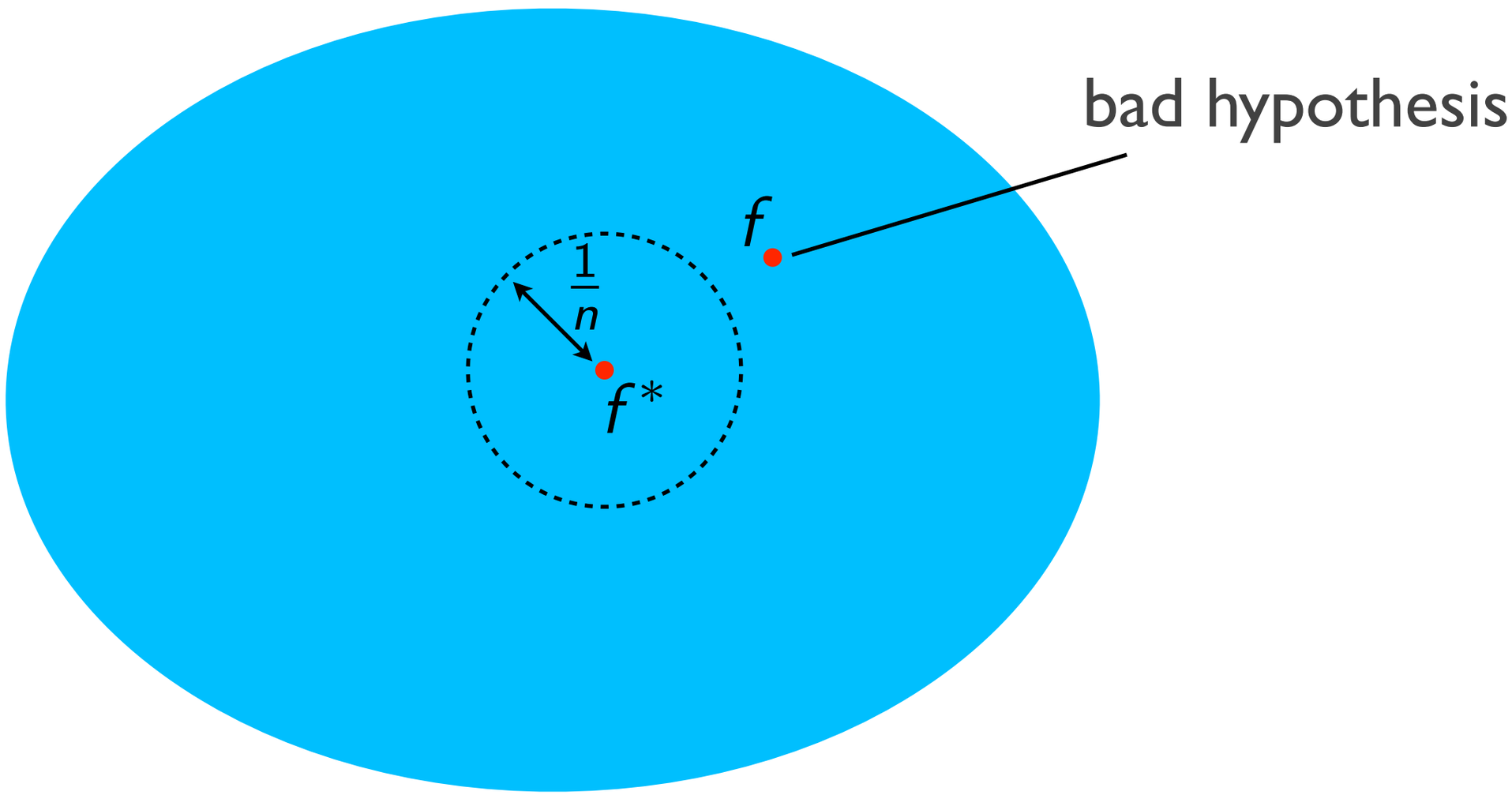
Avoiding Bad Hypotheses

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f(X_j)) \leq \frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f^*(X_j))\right)$$



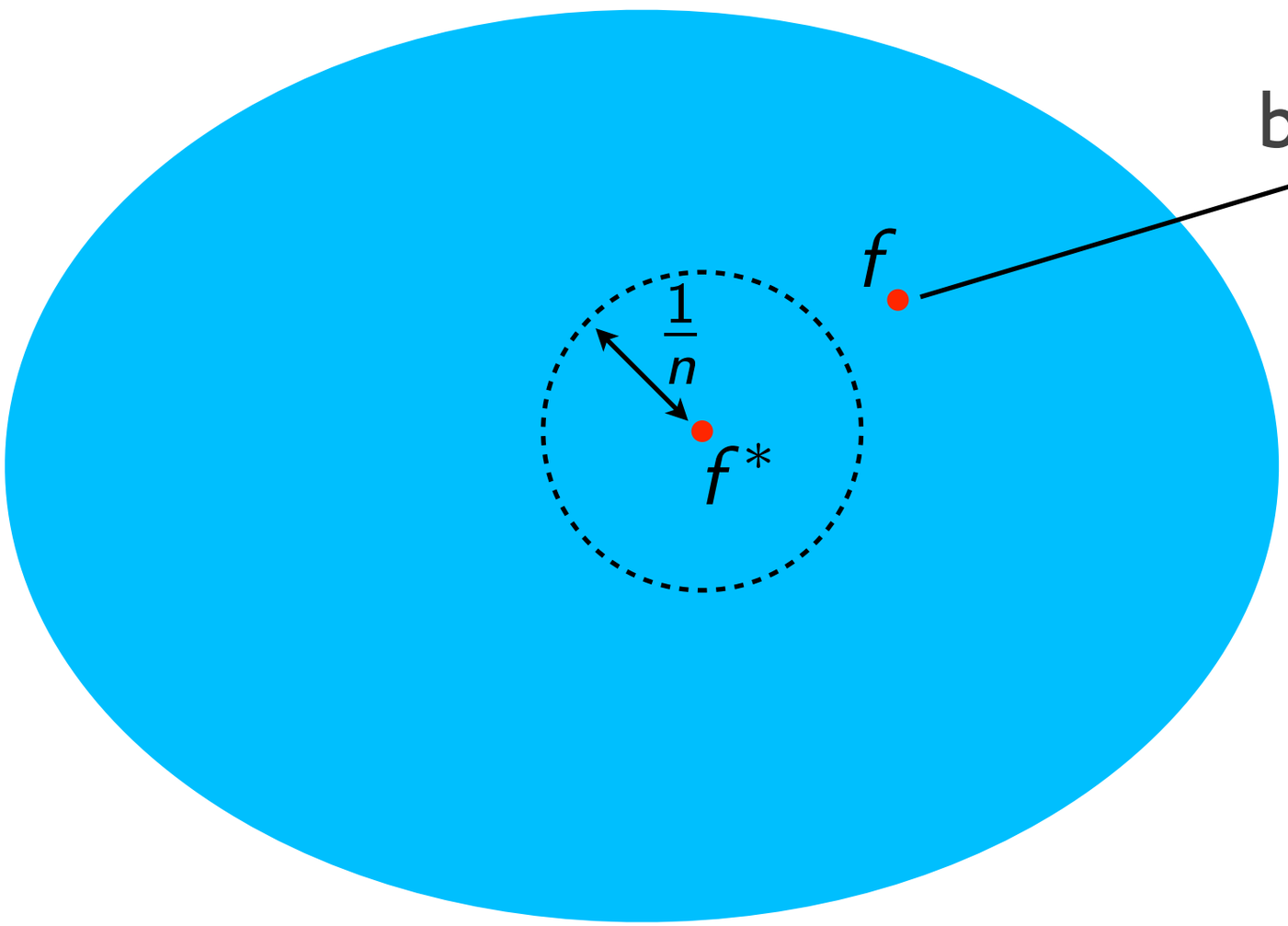
Avoiding Bad Hypotheses

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f(X_j)) \leq \frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f^*(X_j))\right)$$

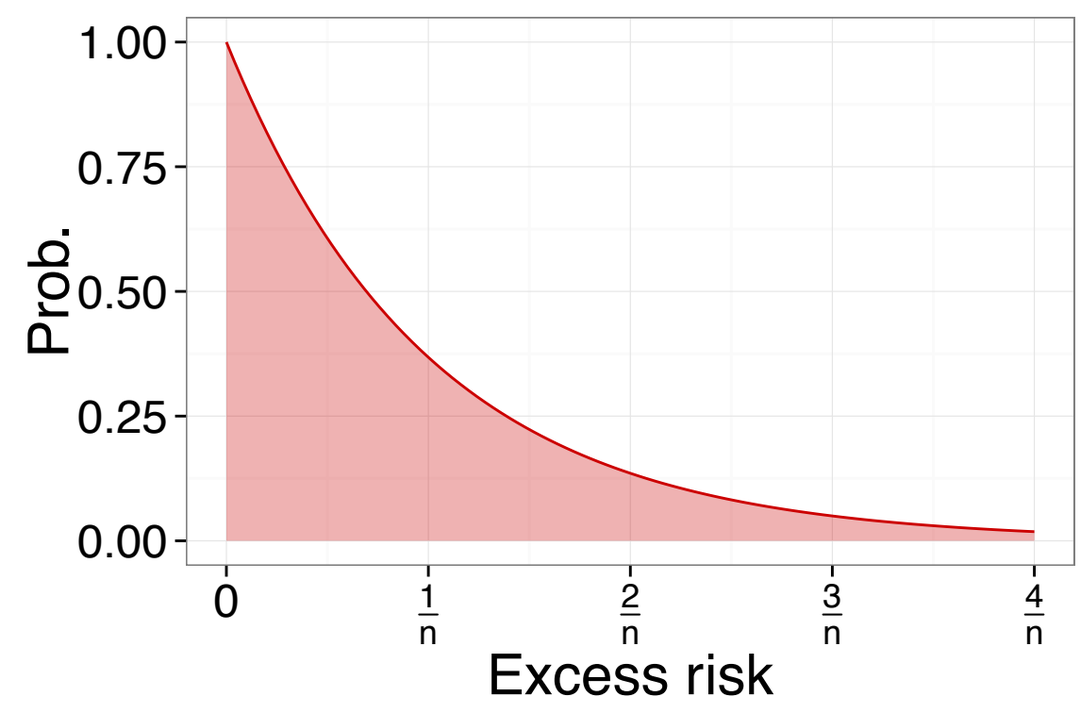


Avoiding Bad Hypotheses

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f(X_j)) \leq \frac{1}{n}\sum_{j=1}^{n}(\ell(Y_j, f^*(X_j))\right)$$



bad hypothesis



Cramér-Chernoff

$$\Pr(S \geq t) \leq \frac{\mathrm{E}\, S}{t}$$

Cramér-Chernoff

$$\Pr(S \geq t) \leq \frac{\mathrm{E}\, S}{t}$$

$$\Pr(S \geq 0) \leq \exp\big(\log \mathrm{E} \exp(\eta S)\big)$$

Cramér-Chernoff

$$\Pr(S \geq t) \leq \frac{\mathrm{E}\, S}{t}$$

$\Lambda_S(\eta)$ cumulant generating function

$$\Pr(S \geq 0) \leq \exp\big(\log \mathrm{E} \exp(\eta S)\big)$$

Cramér-Chernoff

$$\Pr(S \geq t) \leq \frac{\mathrm{E}\, S}{t}$$

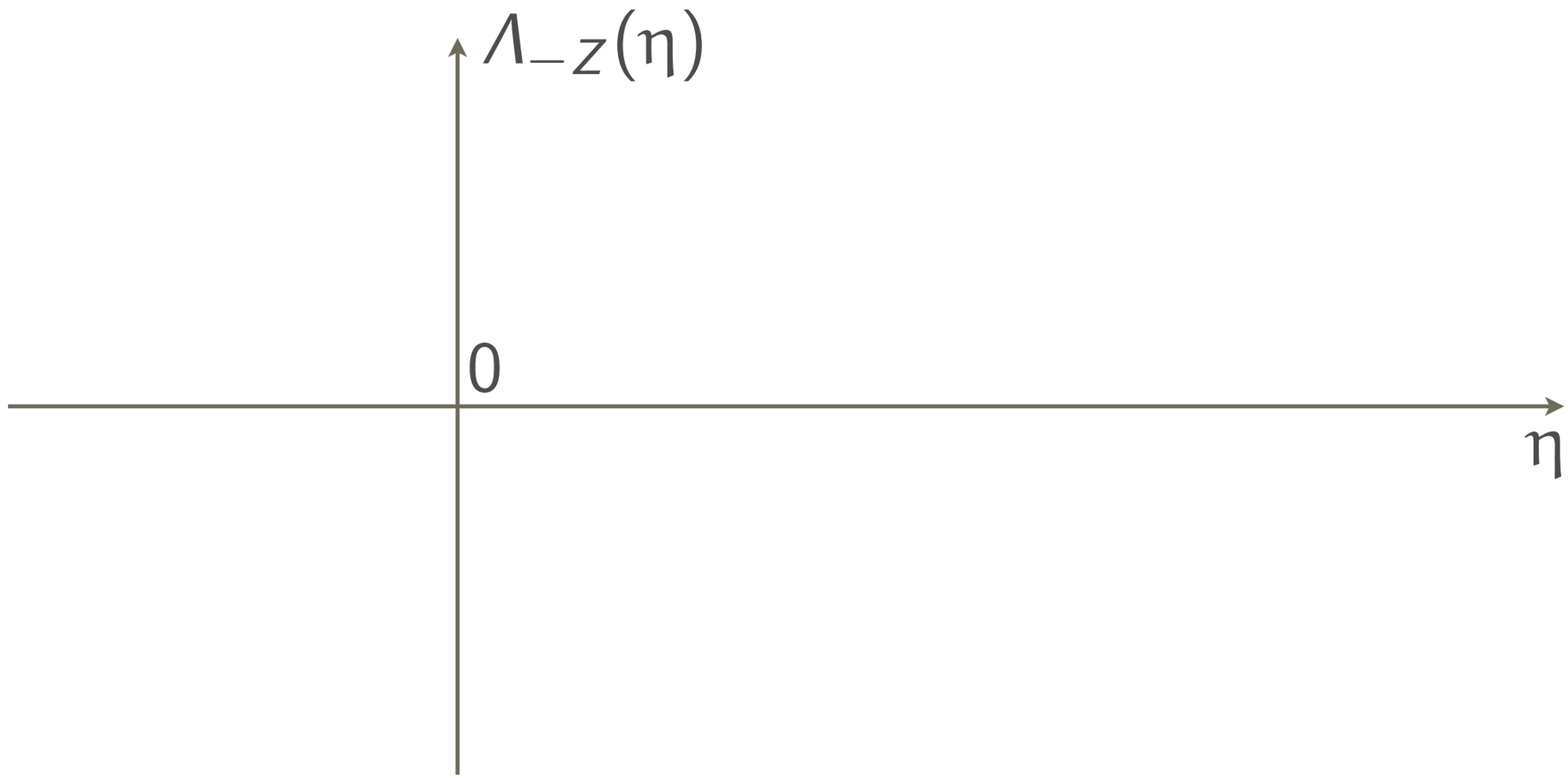
$\Lambda_S(\eta)$ cumulant generating function

$$\Pr(S \geq 0) \leq \exp\big(\log \mathrm{E} \exp(\eta S)\big)$$

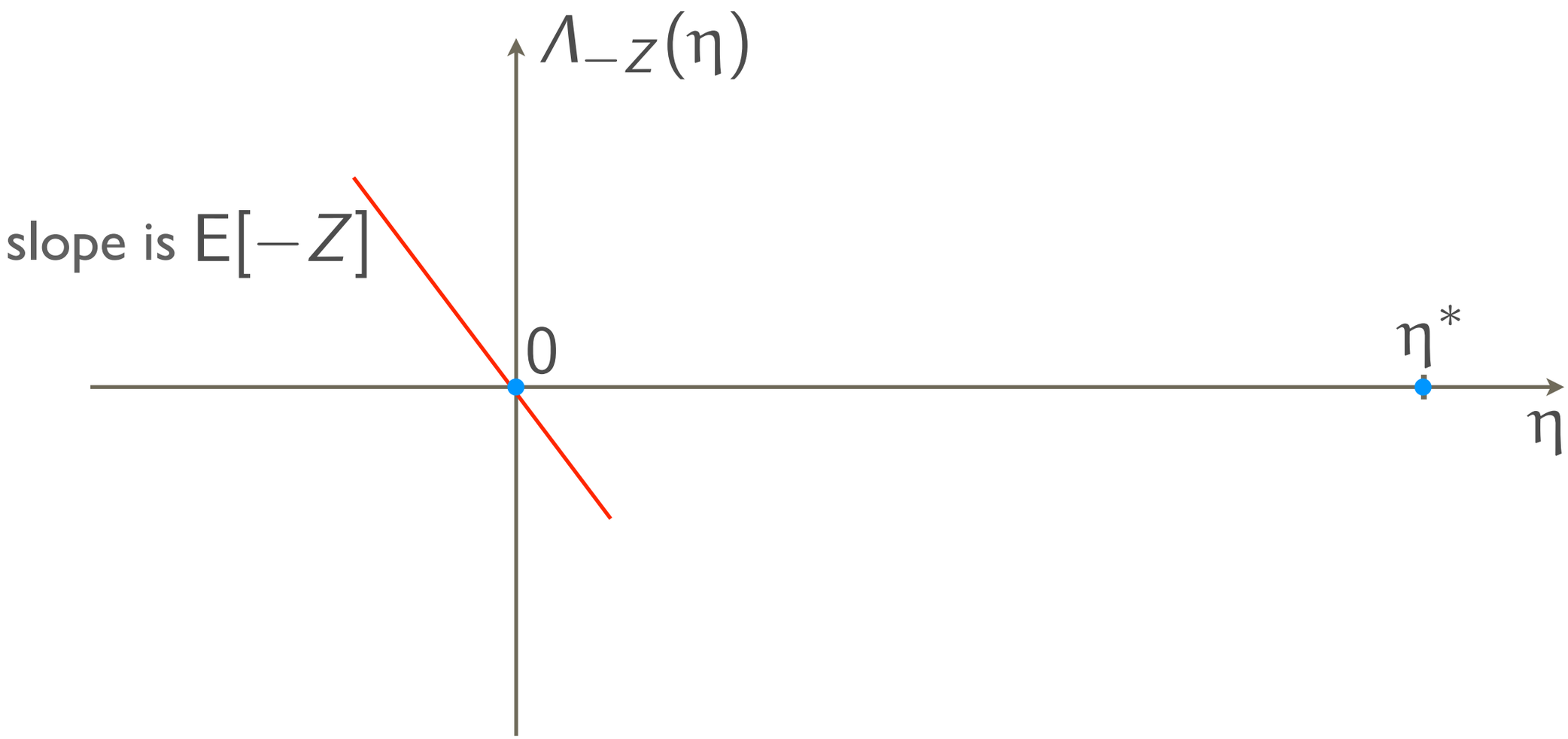
S = average of iid negative excess losses

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\Big[\ell\big(Y_j, f(X_j)\big) - \big(\ell\big(Y_j, f^*(X_j)\big)\Big] \leq 0\right) \leq \exp(n\Lambda_{-Z}(\eta))$$

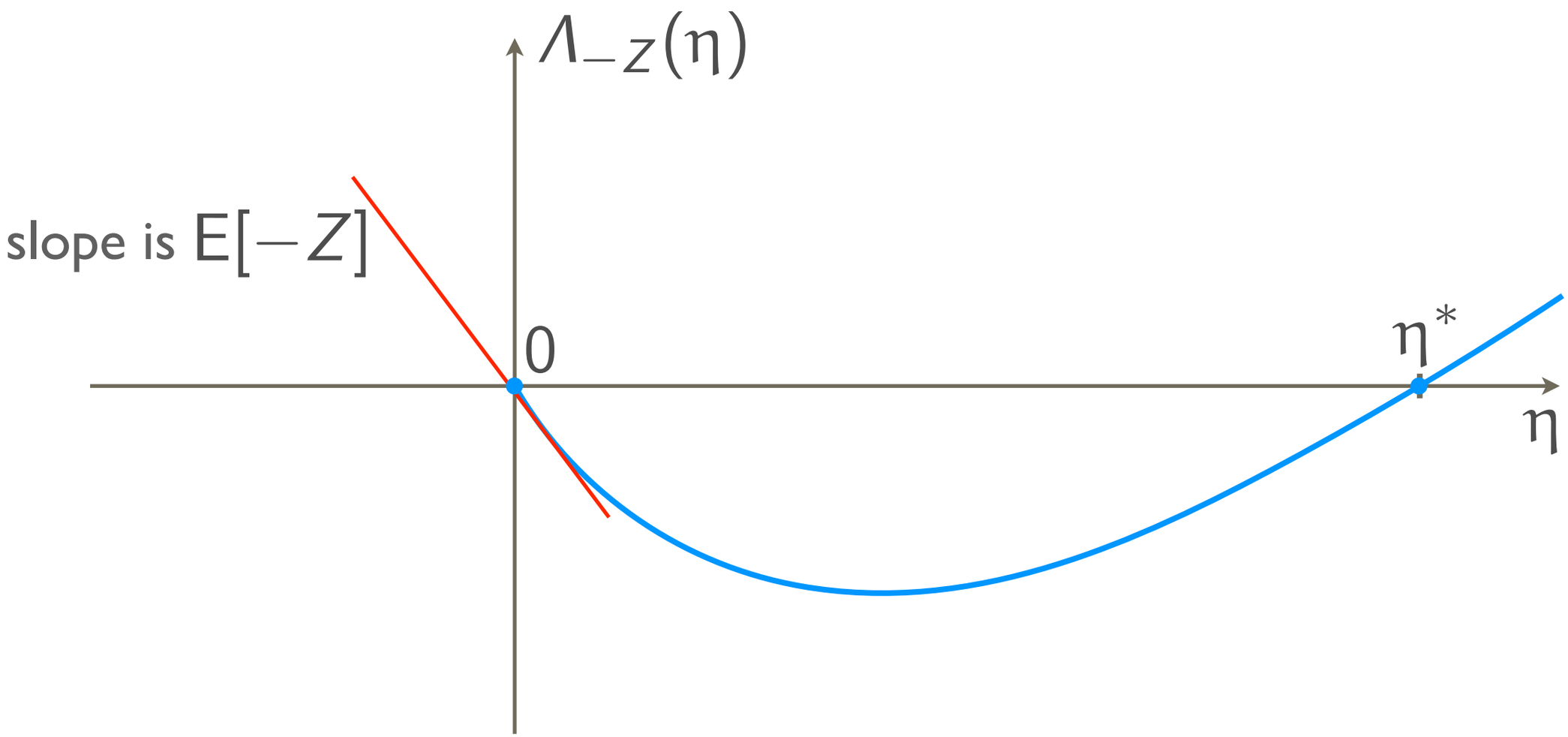
The key picture



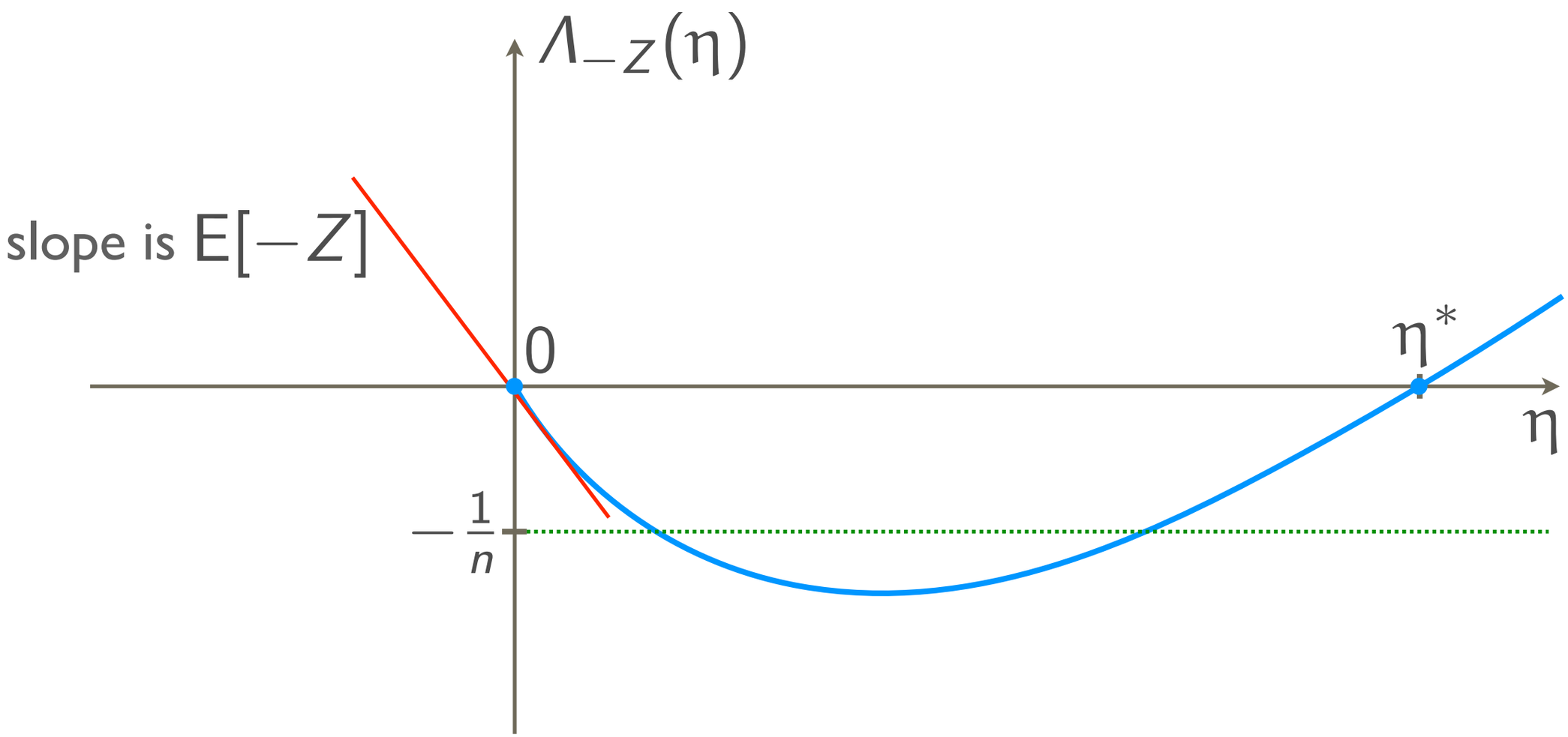
The key picture



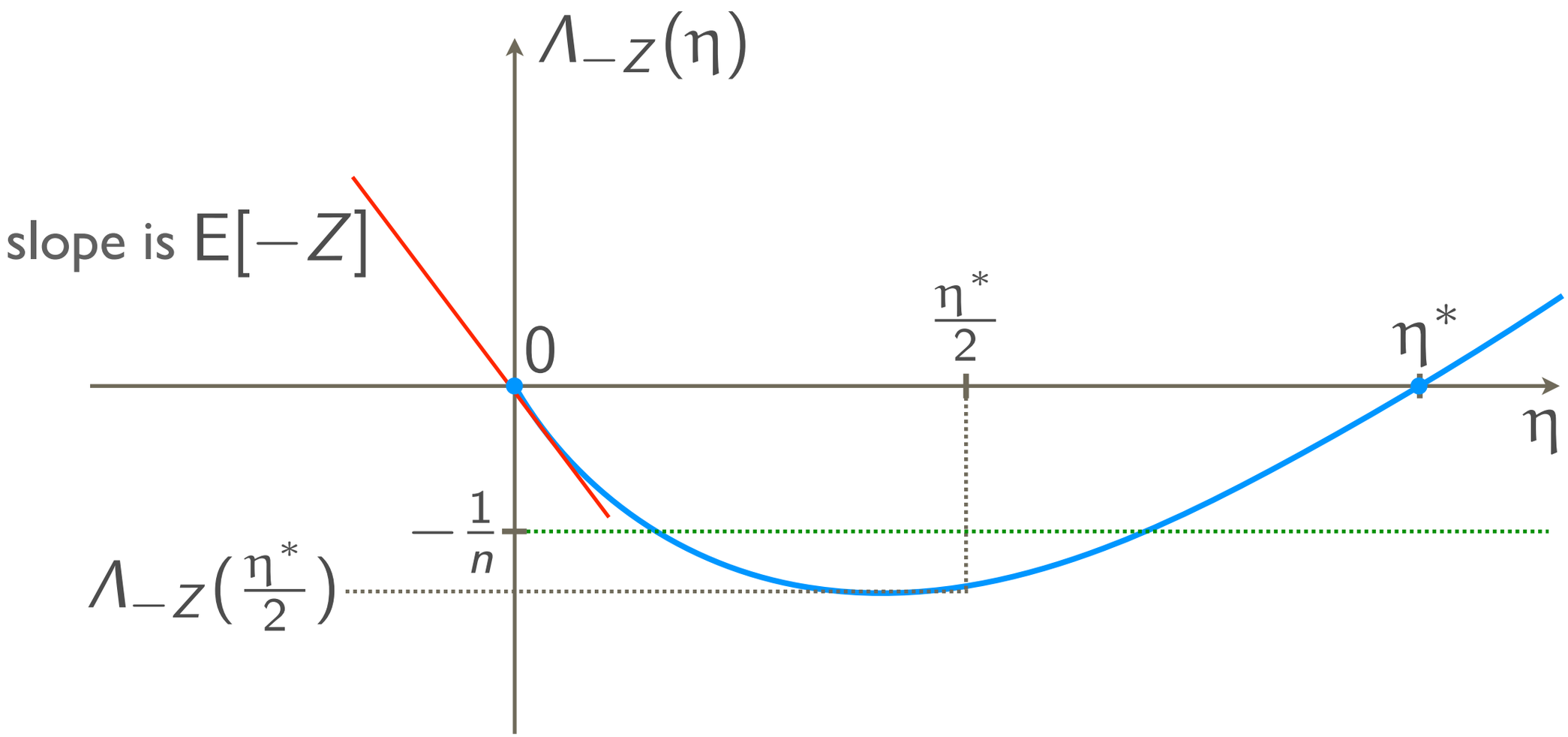
The key picture



The key picture



The key picture



The worst distribution

subject to

Among all distributions on $[-1, 1]$ with mean $-\mu$ $\quad \mathsf{E}_{\mathsf{P}}\, Z = -\mu$

and whose CGF satisfies $\Lambda(\eta^*) = 0$, $\quad \mathsf{E}_{\mathsf{P}}\, e^{\eta^* Z} = 1$

what is the largest that $\Lambda(\frac{\eta^*}{2})$ can be? $\quad \sup_{\mathcal{P}([-1,1])} \mathsf{E}_{\mathsf{P}}\, e^{\frac{\eta^*}{2} Z}$

LP and Semi-infinite LP duality

$$
\begin{aligned}
\max_{x \in \mathbb{R}^n} \quad & \langle c, x \rangle \\
\text{s.t.} \quad & \langle A_j, x \rangle = b_j, \; j \in [m] \\
& x \geq 0
\end{aligned}
$$

LP and Semi-infinite LP duality

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \langle c, x \rangle \\ \text{s.t.} \quad & \langle A_j, x \rangle = b_j, \ j \in [m] \\ & x \geq 0 \end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned} \min_{y \in \mathbb{R}^m} \quad & \langle b, y \rangle \\ \text{s.t.} \quad & \langle A^i, y \rangle \geq c_i, \ i \in [n] \end{aligned}$$

LP and Semi-infinite LP duality

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \langle c, x \rangle \\ \text{s.t.} \quad & \langle A_j, x \rangle = b_j, \ j \in [m] \\ & x \geq 0 \end{aligned}$$

$$\begin{aligned} \max_{\mathsf{P}} \quad & \mathsf{E}_{\mathsf{P}}\, h(Z) \\ \text{s.t.} \quad & \mathsf{E}_{\mathsf{P}}\, g_j(Z) = b_j, \ j \in [m] \\ & \mathsf{E}_{\mathsf{P}}\, 1 = 1 \\ & \mathsf{P} \geq 0 \end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned} \min_{y \in \mathbb{R}^m} \quad & \langle b, y \rangle \\ \text{s.t.} \quad & \langle A^i, y \rangle \geq c_i, \ i \in [n] \end{aligned}$$

LP and Semi-infinite LP duality

$$\max_{x \in \mathbb{R}^n} \quad \langle c, x \rangle$$
$$\text{s.t.} \quad \langle A_j, x \rangle = b_j, \; j \in [m]$$
$$x \geq 0$$

$$\Updownarrow$$

$$\min_{y \in \mathbb{R}^m} \quad \langle b, y \rangle$$
$$\text{s.t.} \quad \langle A^i, y \rangle \geq c_i, \; i \in [n]$$

The General Moment Problem

[Kemperman]

$$\max_{\mathrm{P}} \quad \mathrm{E}_{\mathrm{P}}\, h(Z)$$
$$\text{s.t.} \quad \mathrm{E}_{\mathrm{P}}\, g_j(Z) = b_j, \; j \in [m]$$
$$\mathrm{E}_{\mathrm{P}}\, 1 = 1$$
$$\mathrm{P} \geq 0$$

LP and Semi-infinite LP duality

$$\max_{x \in \mathbb{R}^n} \quad \langle c, x \rangle$$
$$\text{s.t.} \quad \langle A_j, x \rangle = b_j, \; j \in [m]$$
$$x \geq 0$$

$$\Updownarrow$$

$$\min_{y \in \mathbb{R}^m} \quad \langle b, y \rangle$$
$$\text{s.t.} \quad \langle A^i, y \rangle \geq c_i, \; i \in [n]$$

The General Moment Problem

[Kemperman]

$$\max_{\mathsf{P}} \quad \mathsf{E}_{\mathsf{P}}\, h(Z)$$
$$\text{s.t.} \quad \mathsf{E}_{\mathsf{P}}\, g_j(Z) = b_j, \; j \in [m]$$
$$\mathsf{E}_{\mathsf{P}}\, 1 = 1$$
$$\mathsf{P} \geq 0$$

$$\Updownarrow$$

$$\min_{y \in \mathbb{R}^m} \quad y_0 + \langle b, y_{1:m} \rangle$$
$$\text{s.t.} \quad y_0 + \sum_{j=1}^{m} y_j g_j(t) \geq h(t), \; t \in [-1, 1]$$

Stochastic mixability concentration

$$\sup_{P \in \mathcal{P}([-1,1])} \quad \mathsf{E}_{\mathsf{P}}\, e^{\frac{\eta^*}{2} Z}$$

$$\text{s.t.} \quad \mathsf{E}_{\mathsf{P}}\, Z = -\mu$$

$$\mathsf{E}_{\mathsf{P}}\, e^{\eta^* Z} = 1$$

$$\mathsf{E}_{\mathsf{P}}\, 1 = 1$$

$$\mathsf{P} \geq 0$$

$$\Updownarrow$$

$$\min_{y \in \mathbb{R}^m} \quad y_0 - \mu y_1 + y_2$$

$$\text{s.t.} \quad y_0 + y_1 t + y_2 e^{\eta^* t} \geq e^{\frac{\eta^*}{2} t}, \;\; t \in [-1, 1]$$

Stochastic mixability concentration

$$\sup_{P \in \mathcal{P}([-1,1])} \quad \mathsf{E}_{\mathsf{P}}\, e^{\frac{\eta^*}{2} Z}$$

$$\text{s.t.} \quad \mathsf{E}_{\mathsf{P}}\, Z = -\mu$$

$$\mathsf{E}_{\mathsf{P}}\, e^{\eta^* Z} = 1$$

$$\mathsf{E}_{\mathsf{P}}\, 1 = 1$$

$$\mathsf{P} \geq 0$$

$$\Lambda_{-Z}\left(\frac{\eta^*}{2}\right) \leq -0.18\mu(\eta^* \wedge 1)$$

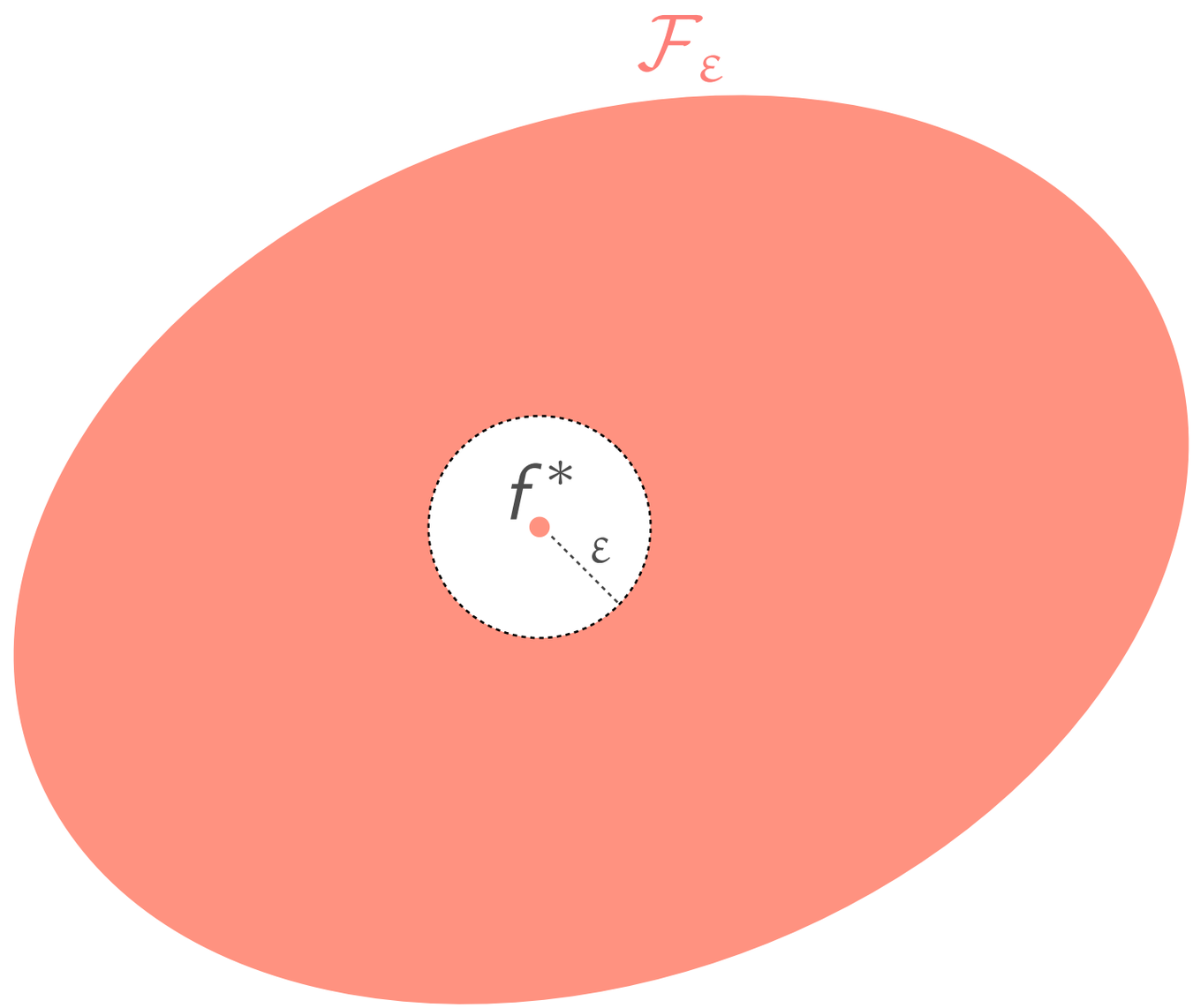
$$\Updownarrow$$

$$\min_{y \in \mathbb{R}^m} \quad y_0 - \mu y_1 + y_2$$

$$\text{s.t.} \quad y_0 + y_1 t + y_2 e^{\eta^* t} \geq e^{\frac{\eta^*}{2} t}, \ t \in [-1, 1]$$

Non-unique minimizers

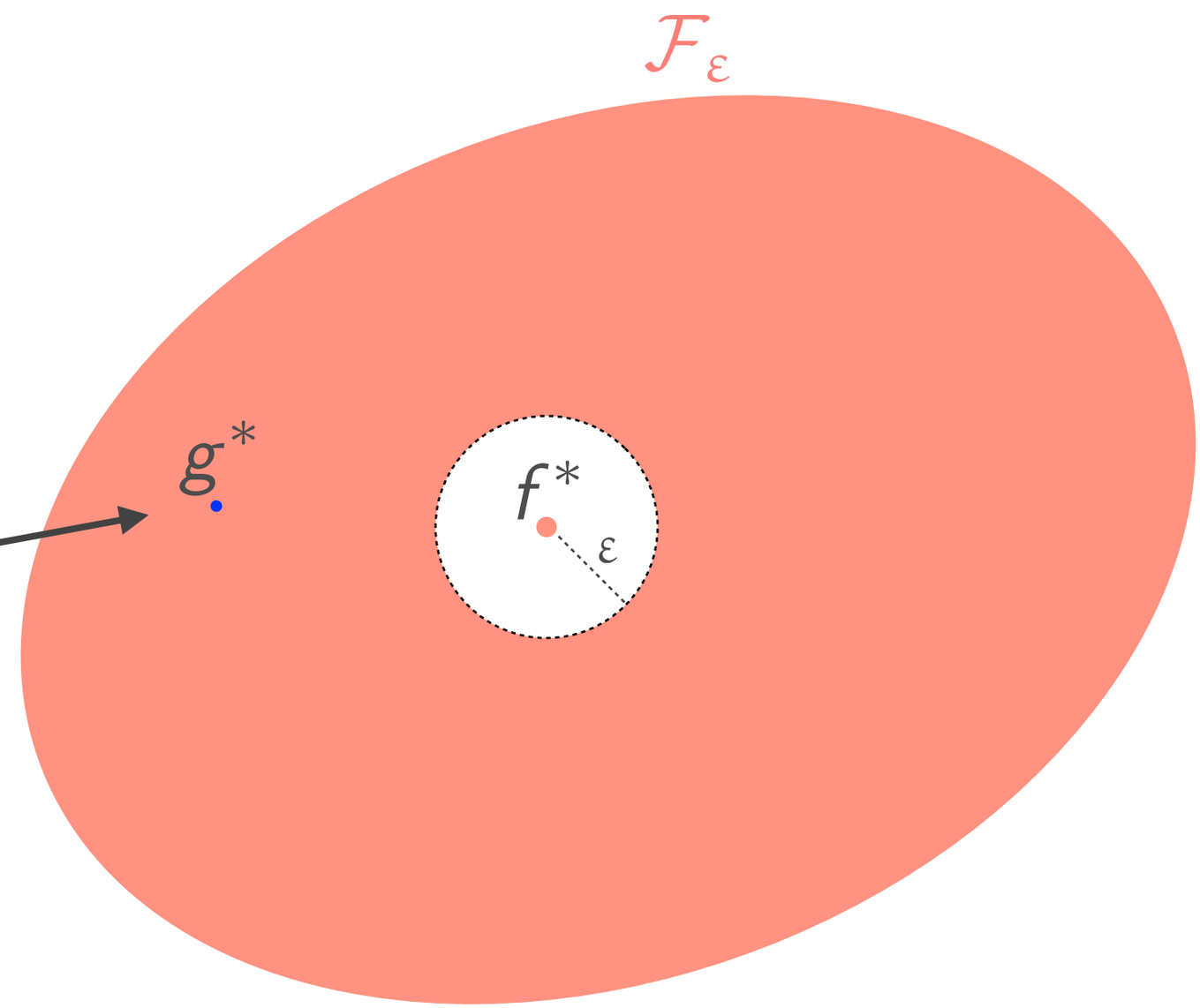
If $(\ell, \mathcal{F}_\varepsilon, P)$ is not stochastically mixable for any $\varepsilon > 0$



Non-unique minimizers

If $(\ell, \mathcal{F}_\varepsilon, P)$ is not stochastically mixable for any $\varepsilon > 0$

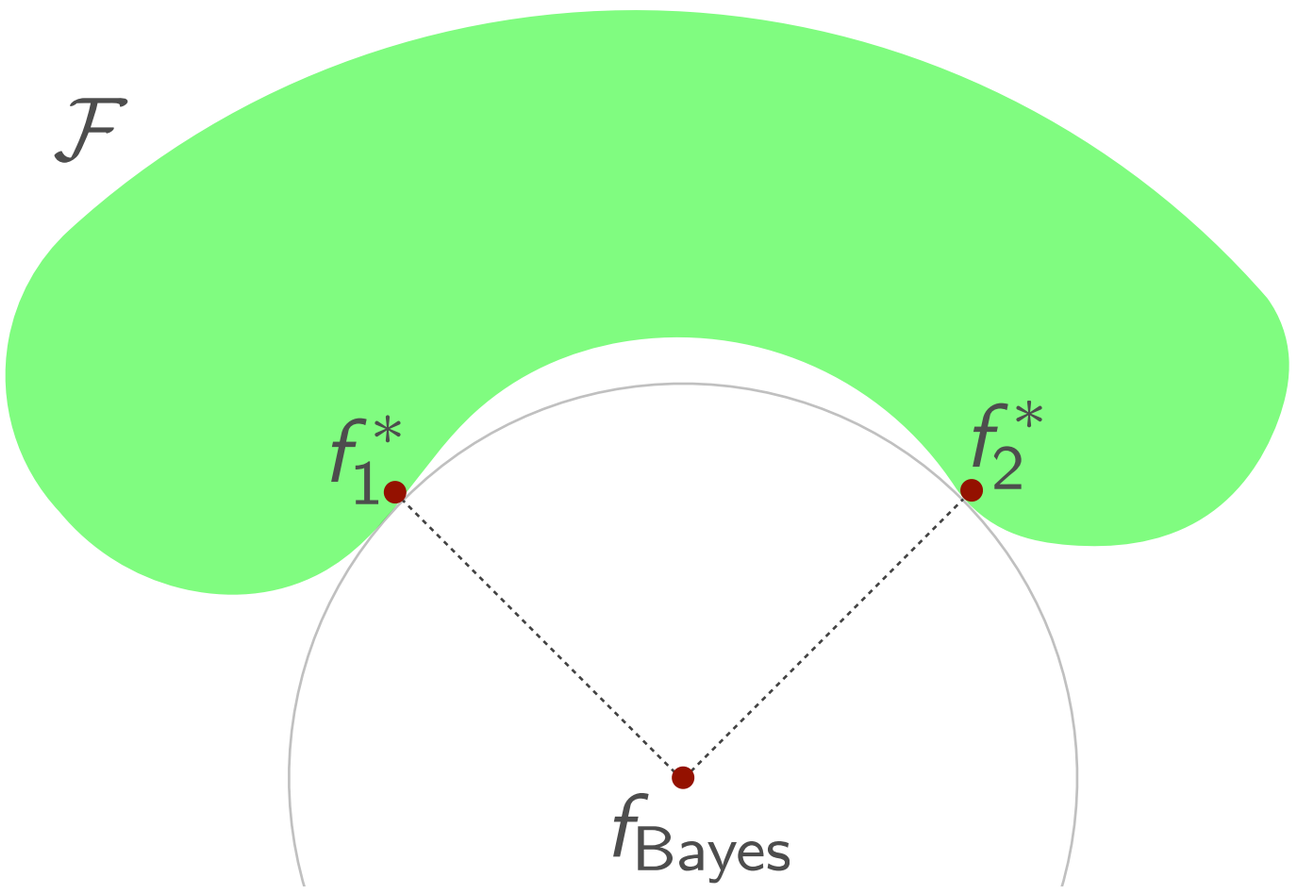
Then



The danger of non-unique minimizers

[Mendelson]

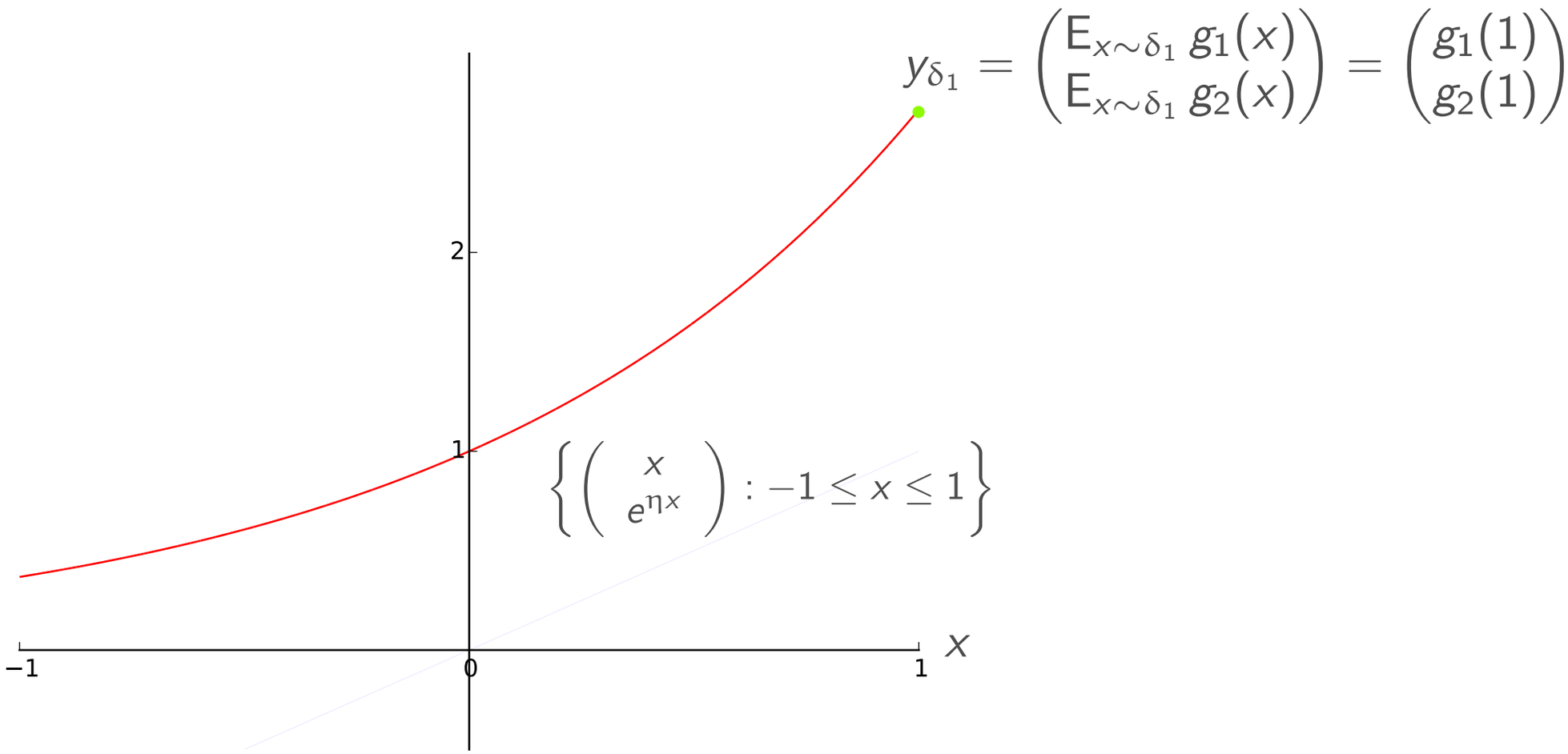
For some n and some target $f^{(n)}$ approaching f_{Bayes} as $n \to \infty$
ERM learns at a slow rate for p-losses



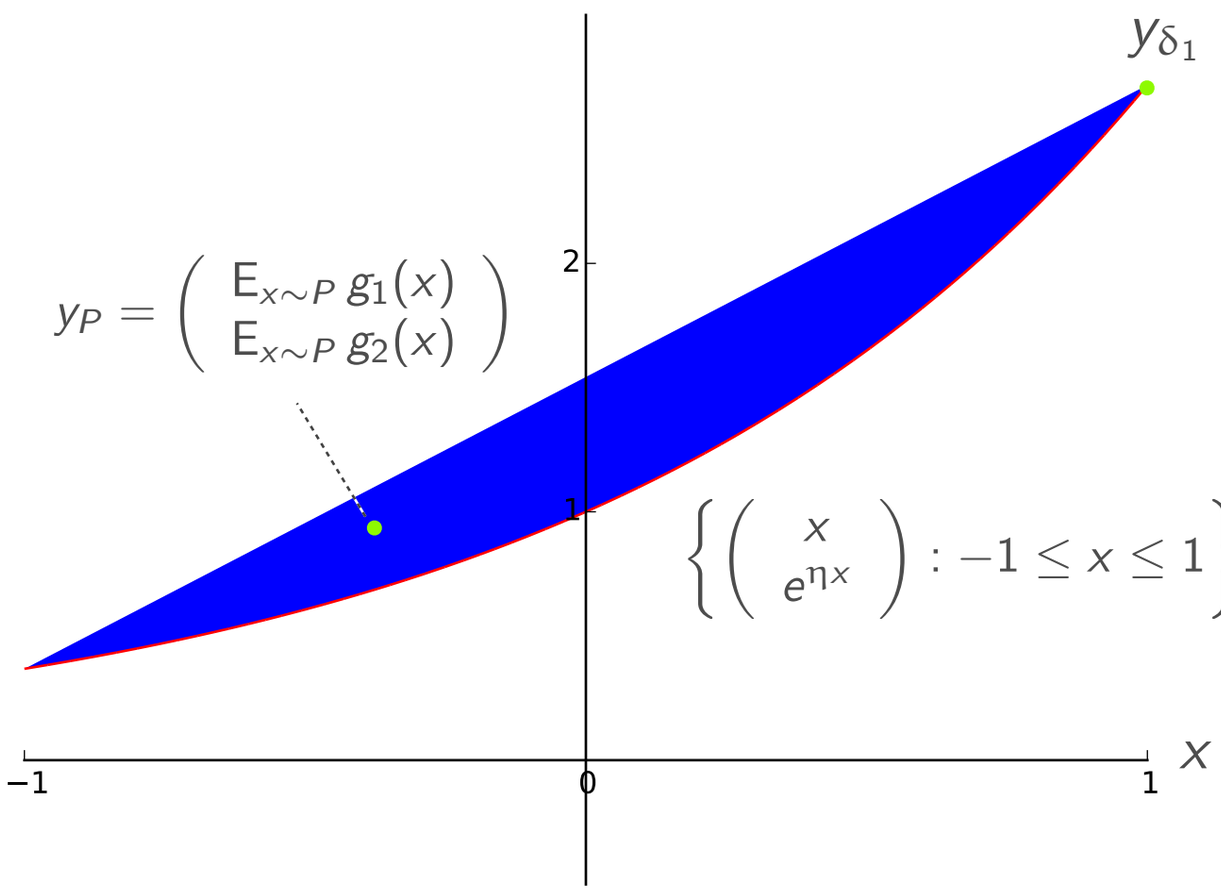
Conclusions

- We strengthened the already strong connection between mixability and stochastic mixability.
- Stochastic mixability implies fast rates for ERM, analogous to how mixability implies constant regret in prediction with expert advice.
- New insights into the fast rates phenomenon through geometry and duality

Connection between excess risk and optimal η

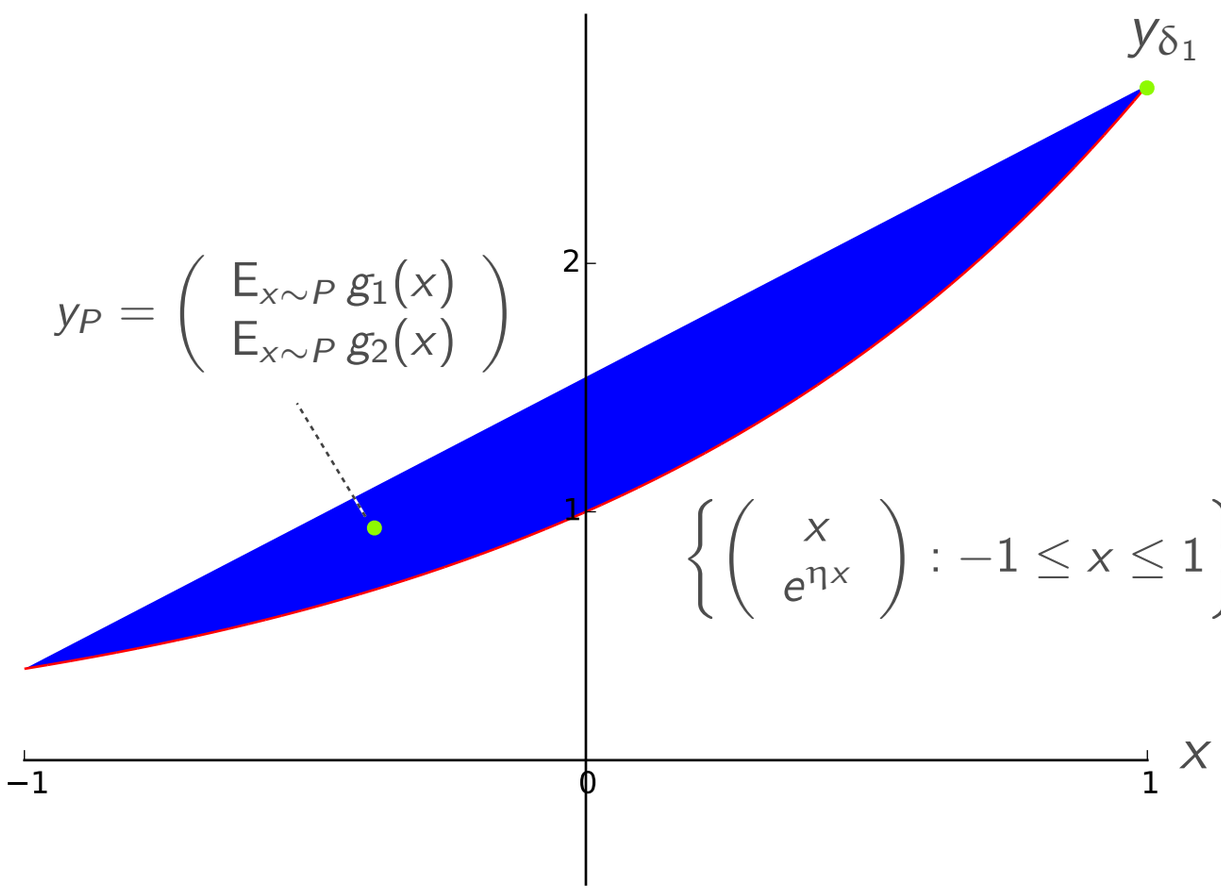


Connection between excess risk and optimal η



$$\Rightarrow \quad \mu \le \frac{\eta}{2}$$

Connection between excess risk and optimal η



$$\Rightarrow \quad \mu \le \frac{\eta}{2}$$

Upshot: for some f, if optimal $\eta \downarrow 0$, then excess risk $\to 0$

Connections to other conditions

[Grünwald, Van Erven, Reid, Williamson]

Stochastic mixability $\longleftrightarrow$ Bernstein condition

$\mathsf{E}\, e^{-\eta(f-f^*)} \leq 1$

bounded losses

$\alpha = 1$

$\mathsf{E}(f-f^*)^2 \leq c\,\mathsf{E}(f-f^*)^\alpha$

$\downarrow$ $f^* = f_{\text{Bayes}} \in \mathcal{F}$

Margin condition

$\downarrow$ $\alpha = 1$

Massart's noise condition

BERNSTEIN IS 2-SIDED,
STOCHASTIC MIXABILITY IS 1-SIDED!