Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds

(presented at COLT2023)

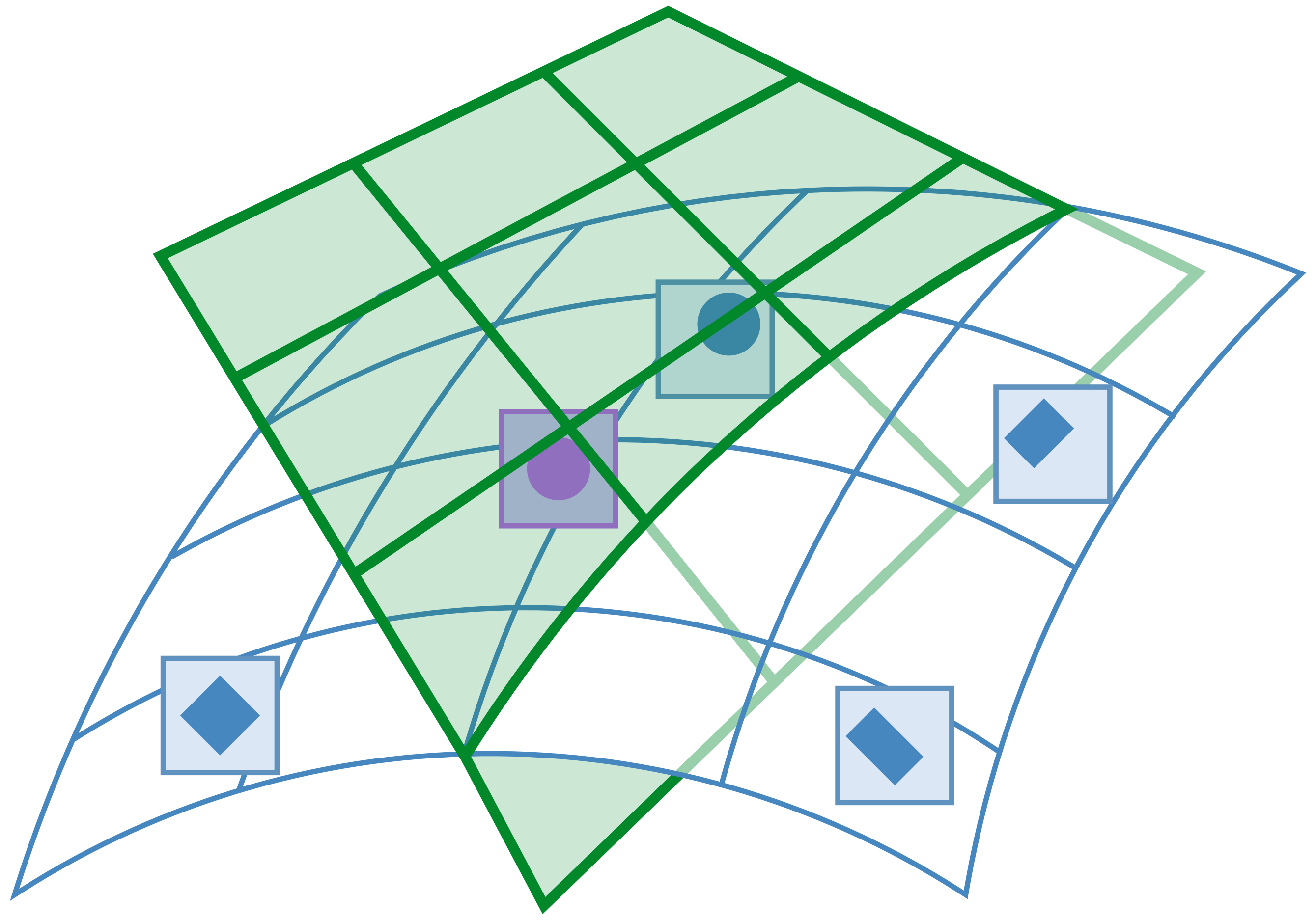
Han Bao

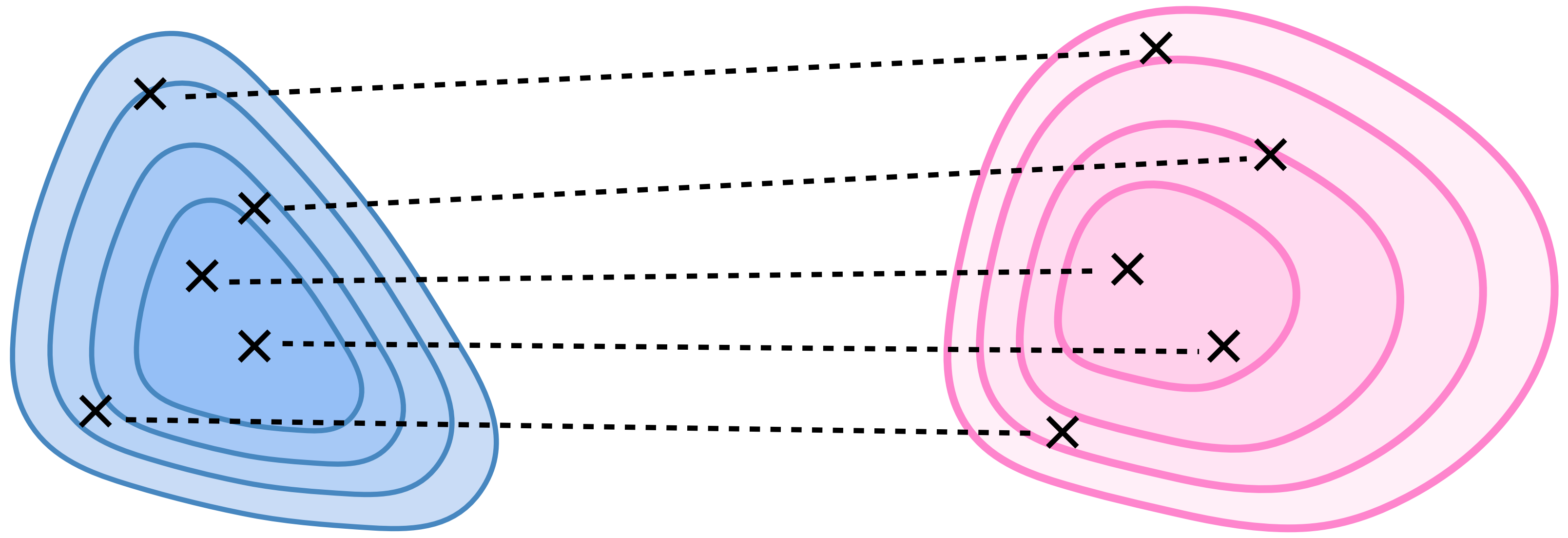
October 13rd, 2023 @ Bristol

KYOTO UNIVERSITY

Short bio | Han Bao

- 2017 April - 2022 March: Graduate student @ The University of Tokyo
- 2022 April - Current: Assistant professor @ Kyoto University (Hakubi center)
- Research keywords:
  Loss function **(today's topic)**, robustness, contrastive learning, optimal transport ⋯





Bandit 🇨🇦

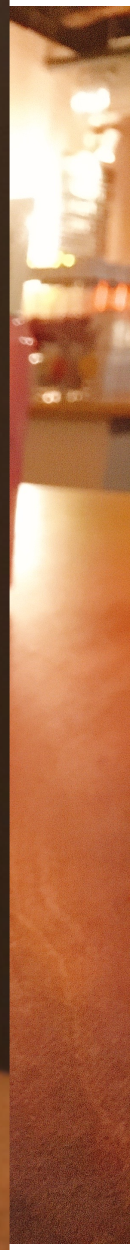
Lervig 🇳🇴

Uchu 🇯🇵

Omnipollo 🇸🇪

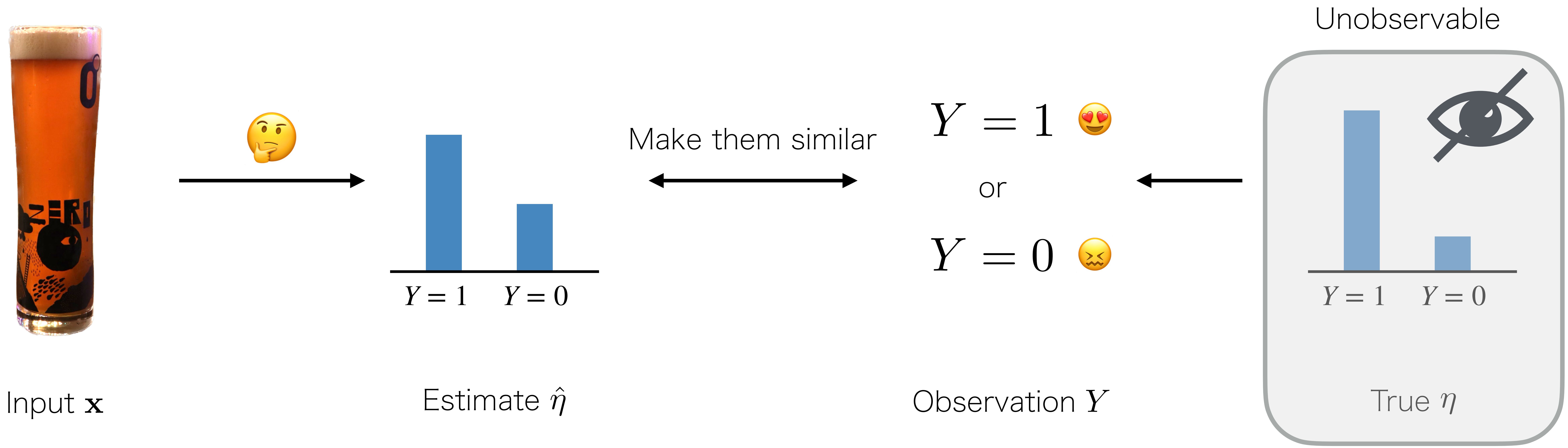
Bandit 🇨🇦

Lervi



Probabilistic prediction

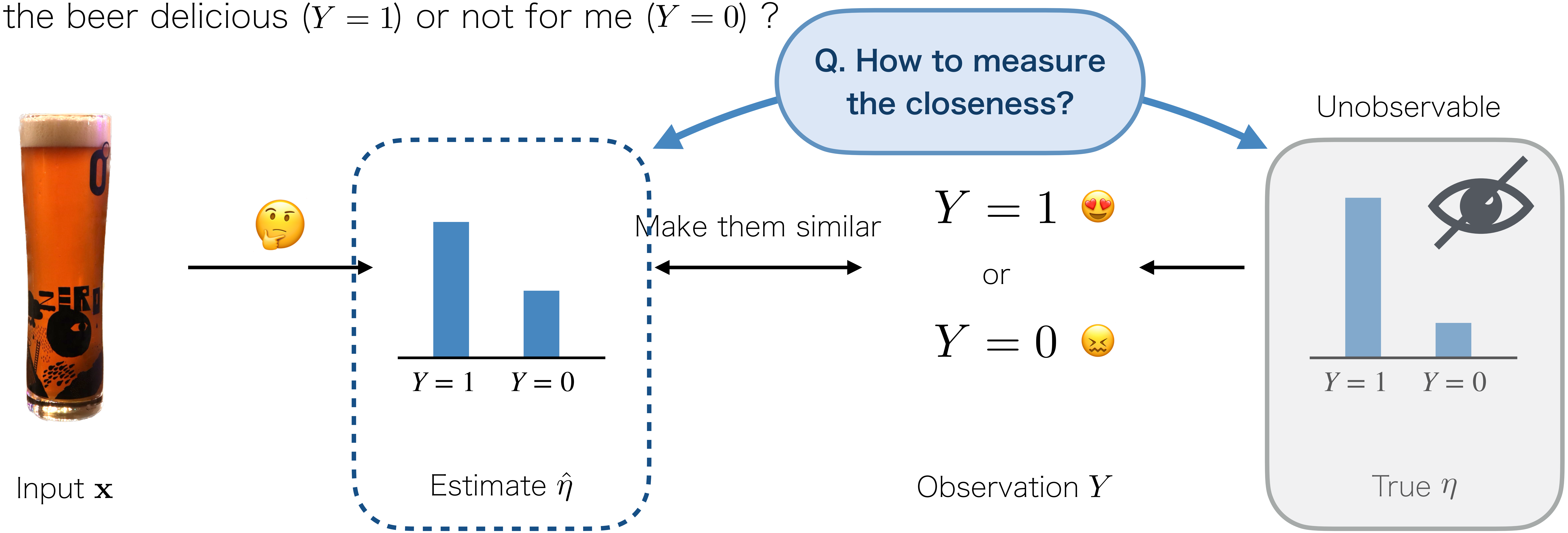
- Is the beer delicious ($Y = 1$) or not for me ($Y = 0$) ?



- We make a decision $\hat{\eta}$ that is as close to true $\eta = \mathbb{P}(Y = 1 \mid \mathbf{x})$ as possible

Probabilistic prediction

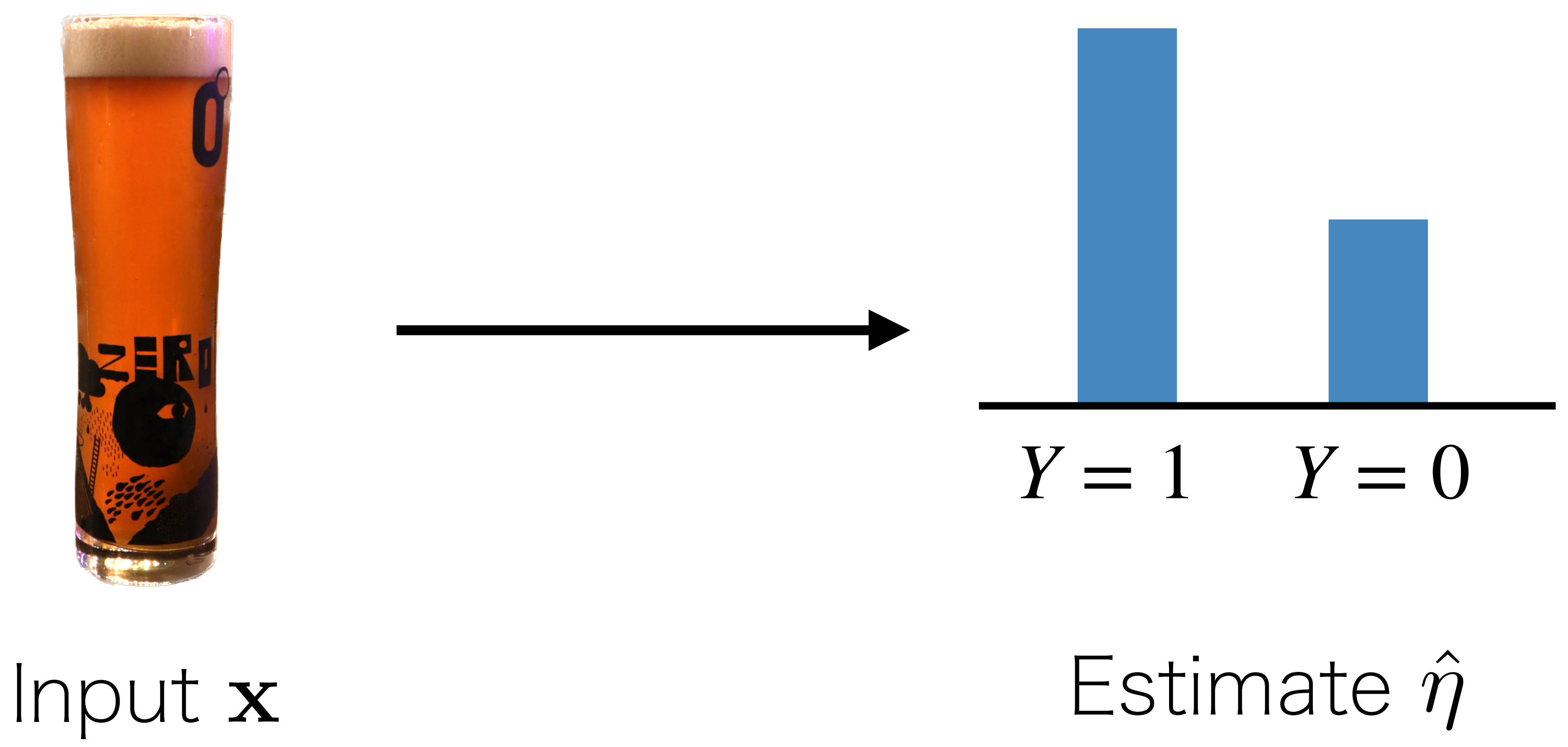
- Is the beer delicious ($Y = 1$) or not for me ($Y = 0$) ?



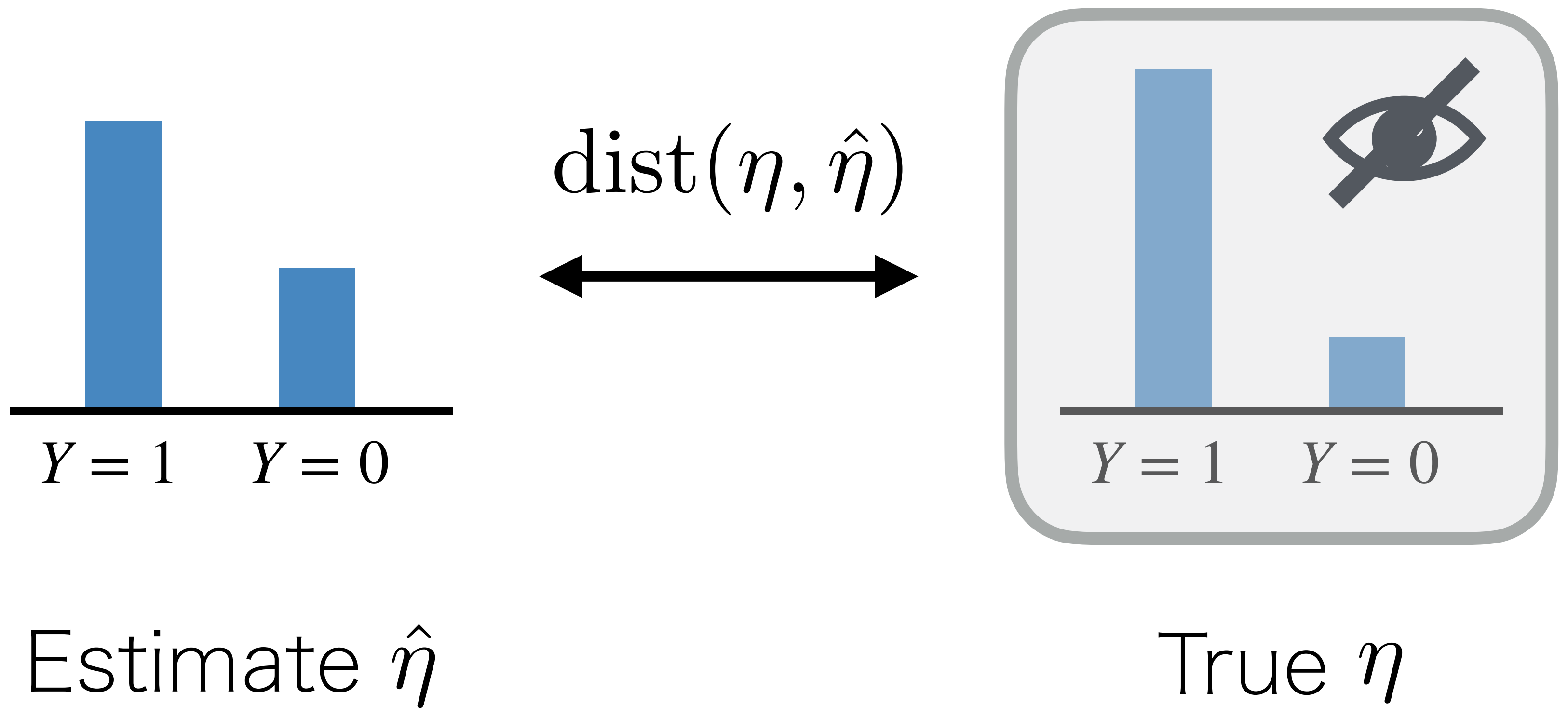
- We make a decision $\hat{\eta}$ that is as close to true $\eta = \mathbb{P}(Y = 1 \mid \mathbf{x})$ as possible

Formulation | Probabilistic prediction

- Input: sample $\{(\mathbf{x}_i, y_i)\}_{i \in [n]}$ with binary outcome $y_i \in \{1, 0\}$
- Goal: to estimate $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x})$
  - Output: $\hat{\eta} : \mathcal{X} \to [0, 1]$ such that $\hat{\eta} \approx \eta$

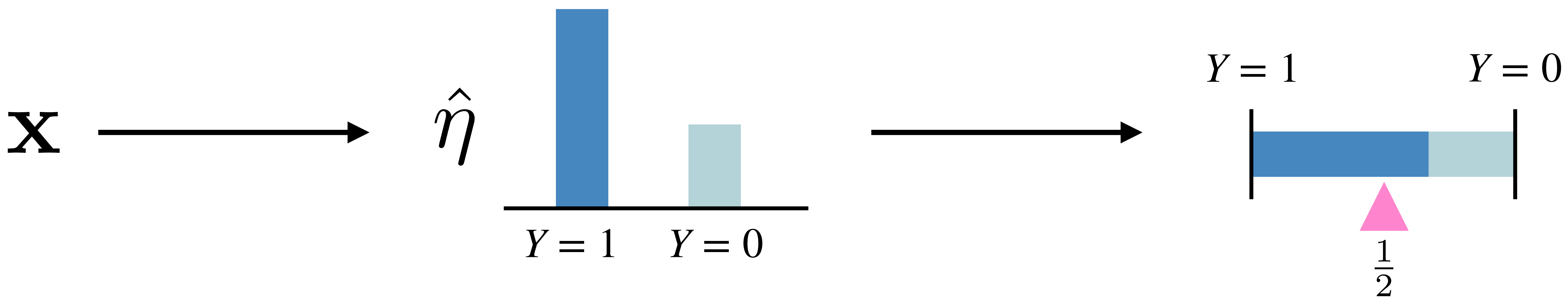


- Challenge: no observation of $\eta(\mathbf{x}_i)$
  - It is important to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y
  - Disclaimer: we do not discuss functional approximation



Probabilistic predictions everywhere

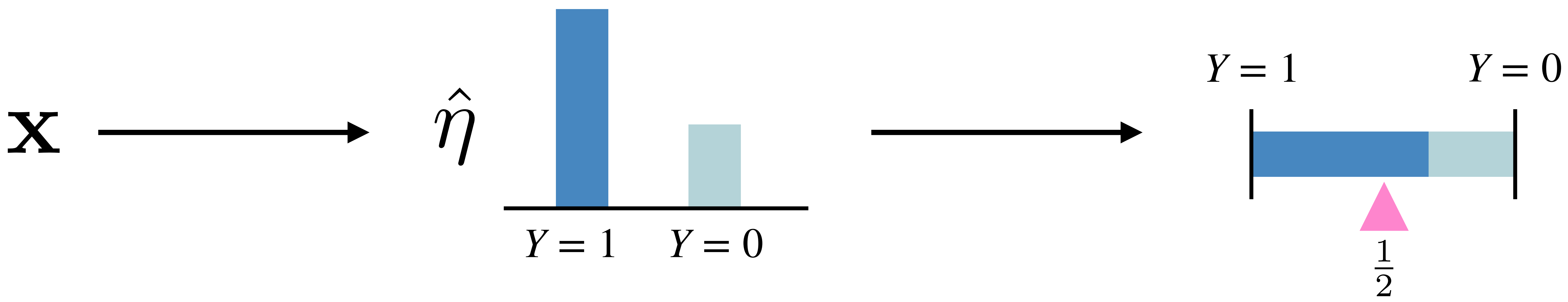
- Having estimated $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x})$, many downstream tasks can be solved
- Case 1 (classification): $f^*(\mathbf{x}) = \mathrm{sign}\big(\eta(\mathbf{x}) - \frac{1}{2}\big)$



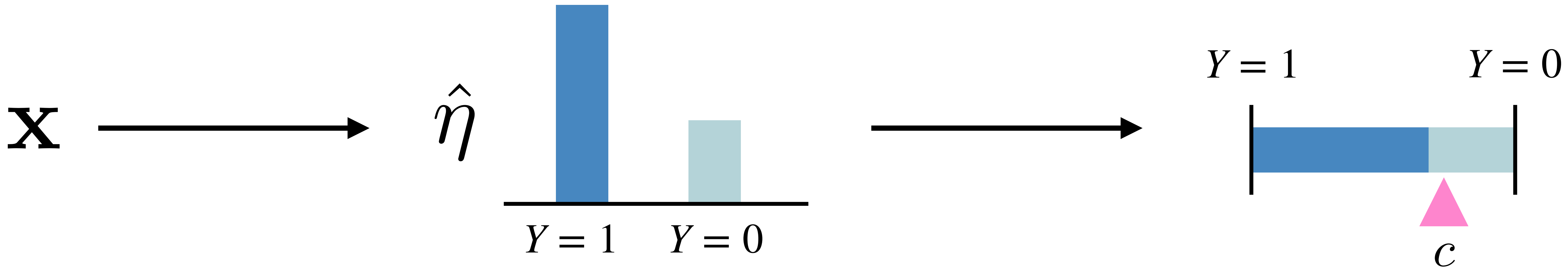
Elkan, Charles. The foundations of cost-sensitive learning. Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001.

Probabilistic predictions everywhere

- Having estimated $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x})$, many downstream tasks can be solved
- Case 1 (classification): $f^*(\mathbf{x}) = \text{sign}\left(\eta(\mathbf{x}) - \frac{1}{2}\right)$



- Case 2 (cost-sensitive classification): $f^*(\mathbf{x}) = \text{sign}(\eta(\mathbf{x}) - c)$ [Elkan 2001]
  - Cost for false positives = c



Elkan, Charles. The foundations of cost-sensitive learning. Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001.

Probabilistic predictions everywhere

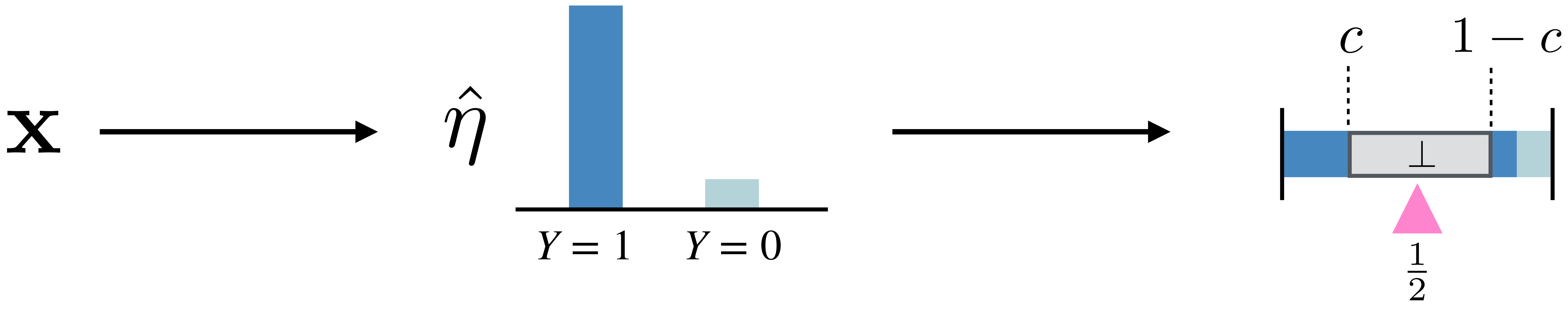
- Having estimated $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x})$, many downstream tasks can be solved
- Case 3 (bipartite ranking): Higher score for positive examples than negative examples
  - $f^*(\mathbf{x}) = \iota(\eta(\mathbf{x}))$ (ι : some monotonic transform)

Chow., C. K. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1):41–46, 1970.

Probabilistic predictions everywhere

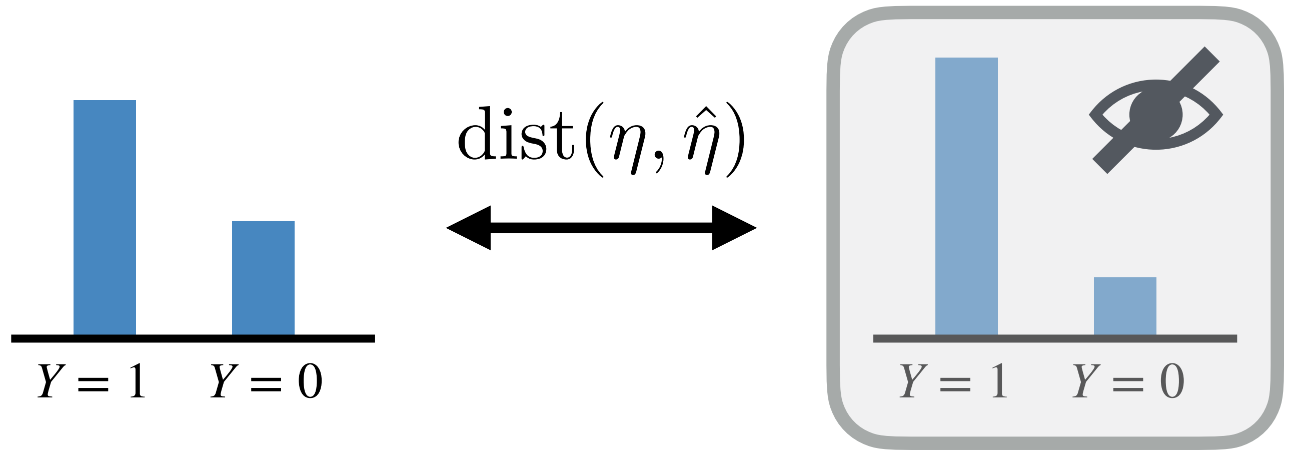
- Having estimated $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{x})$, many downstream tasks can be solved
- Case 3 (bipartite ranking): Higher score for positive examples than negative examples
  - $f^*(\mathbf{x}) = \iota(\eta(\mathbf{x}))$ (ι : some monotonic transform)
- Case 4 (learning with rejection): Agent is allowed to defer decisions to human with cost c
  - $f^*(\mathbf{x}) = \begin{cases} \perp & \text{if } c \leq \eta(\mathbf{x}) \leq 1 - c \\ \operatorname{sign}\left(\eta(\mathbf{x}) - \frac{1}{2}\right) & \text{otherwise} \end{cases}$

[Chow 1970]



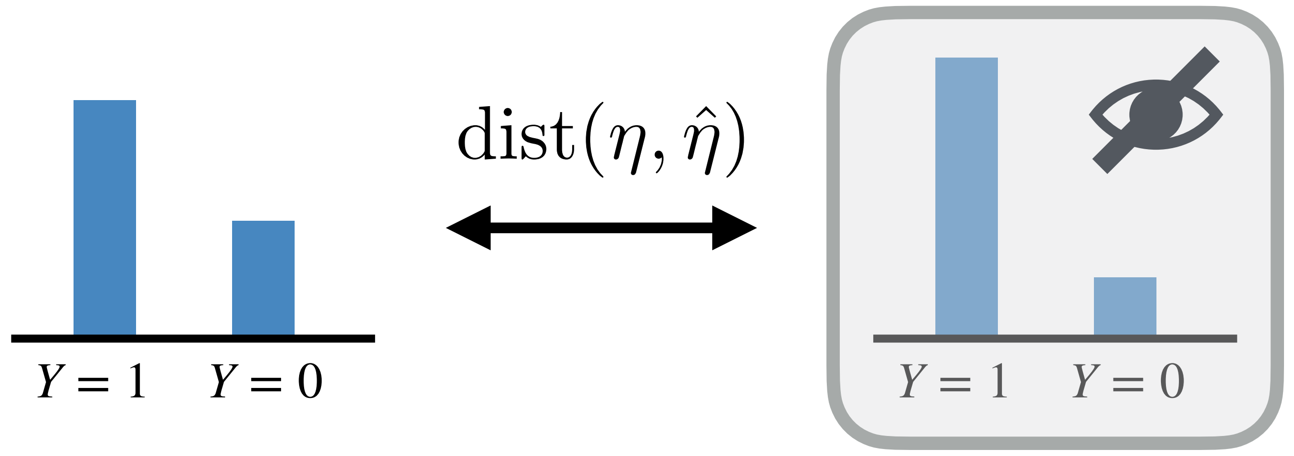
Questions

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?



Questions

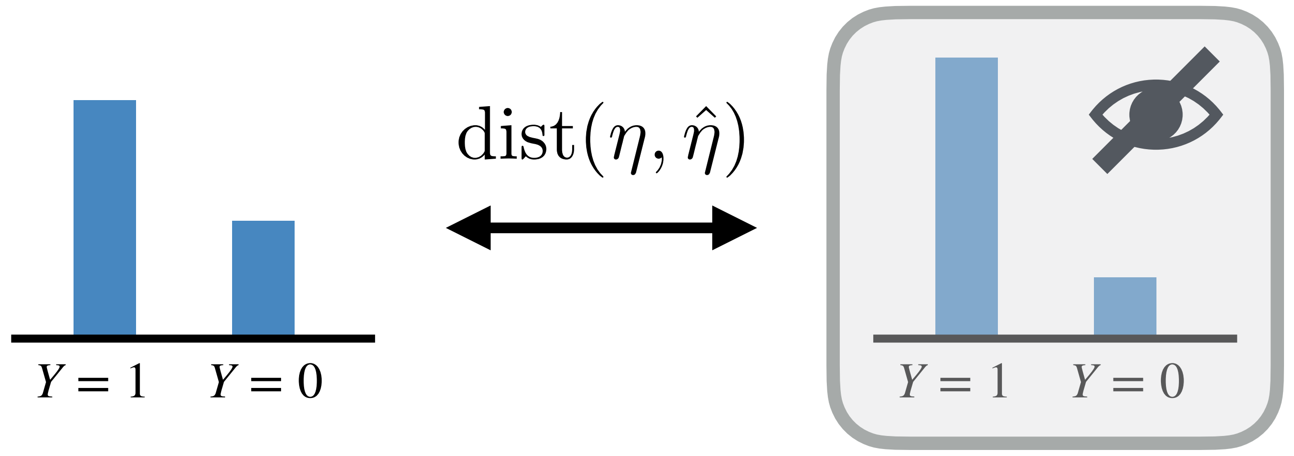
- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?



Proper Losses

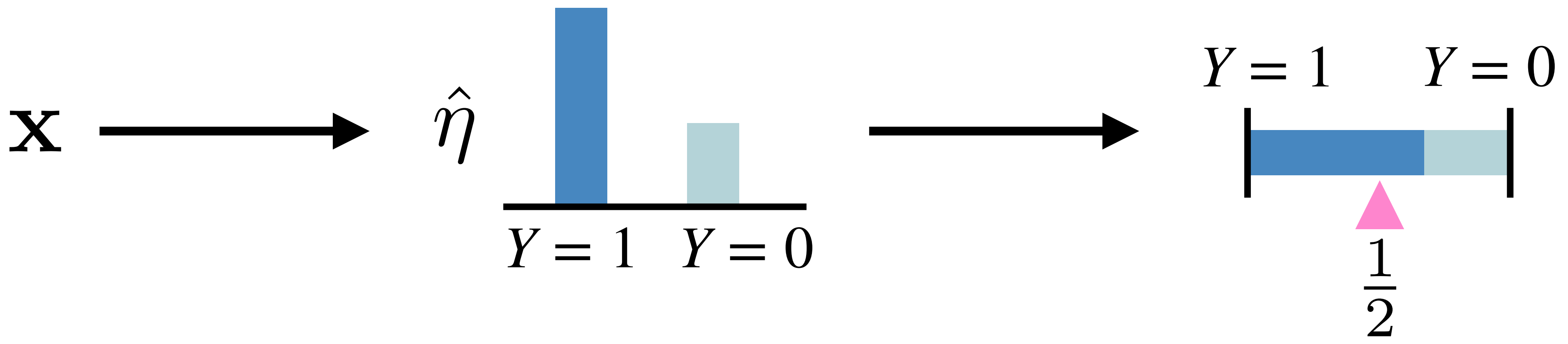
Questions

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?



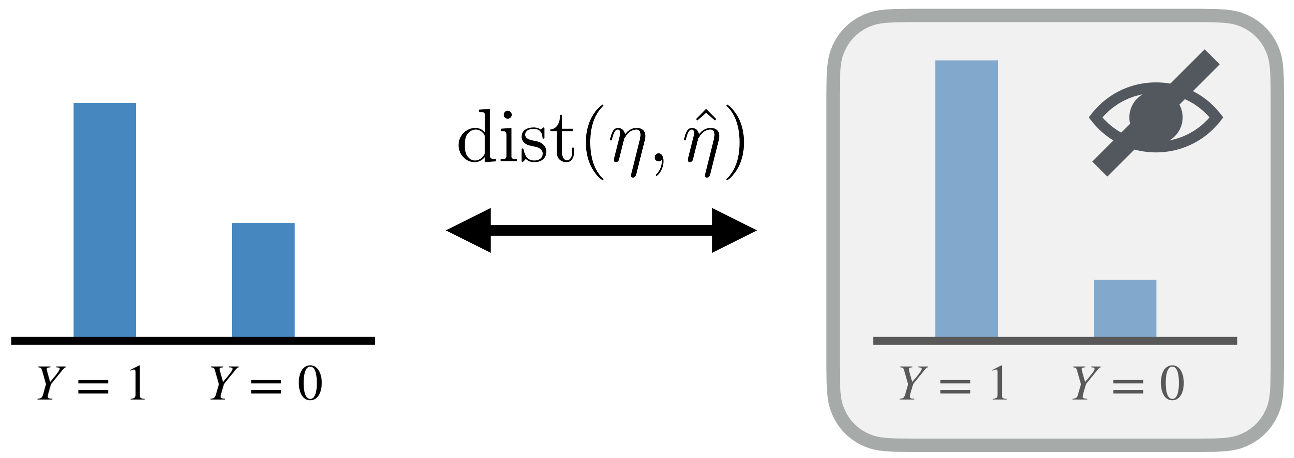
Proper Losses

- How useful is probability estimate $\hat{\eta}$ for a downstream task?



Questions

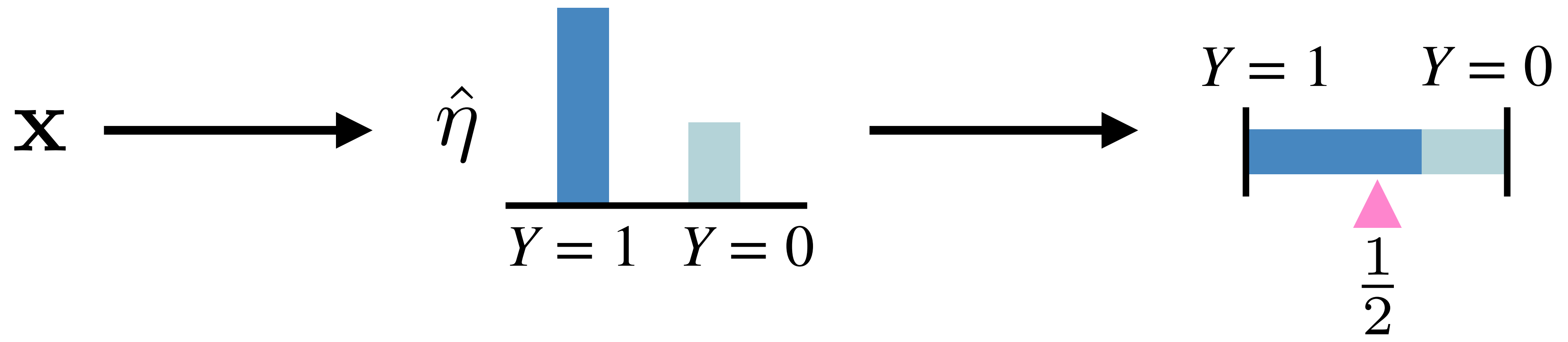
- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?



Proper Losses

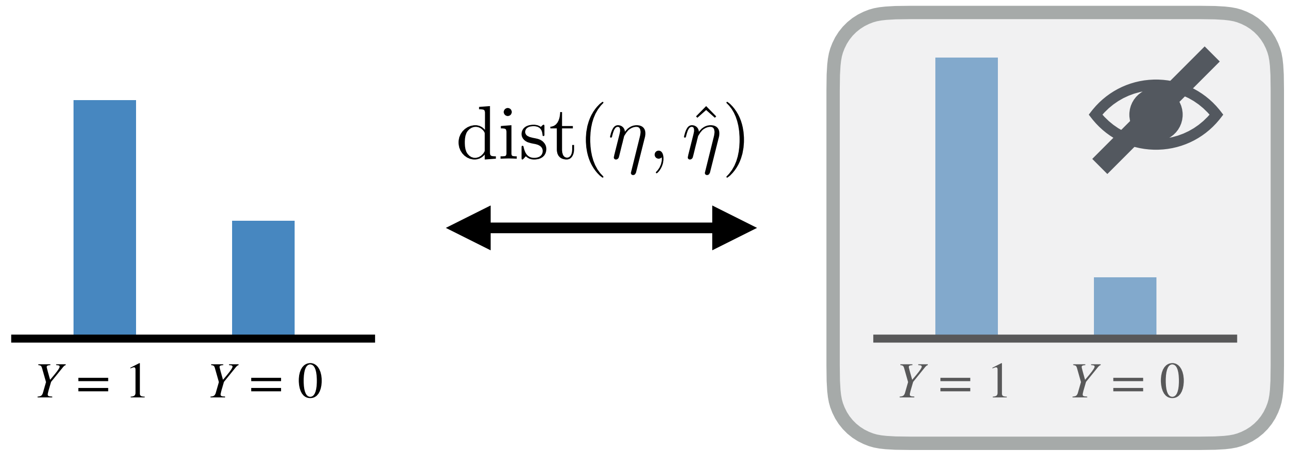
Surrogate Regret Bounds

- How useful is probability estimate $\hat{\eta}$ for a downstream task?



Questions

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?

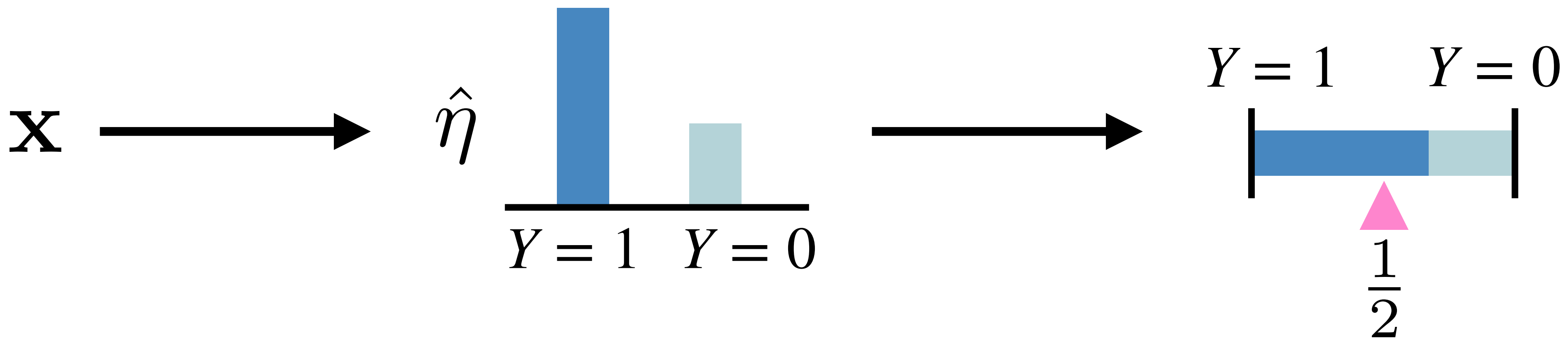


- How to derive surrogate regret bounds?

Proper Losses

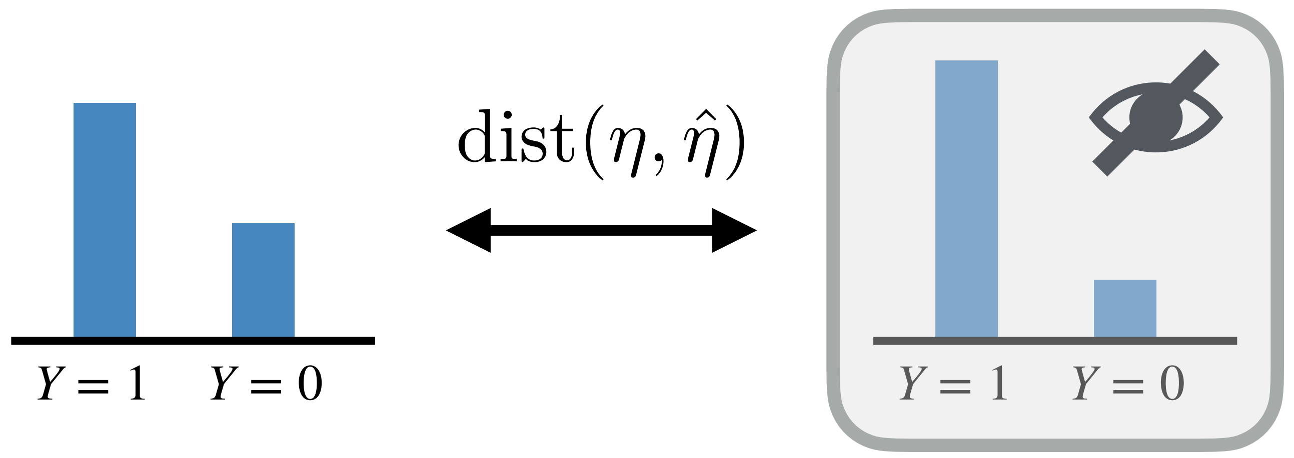
Surrogate Regret Bounds

- How useful is probability estimate $\hat{\eta}$ for a downstream task?



Questions

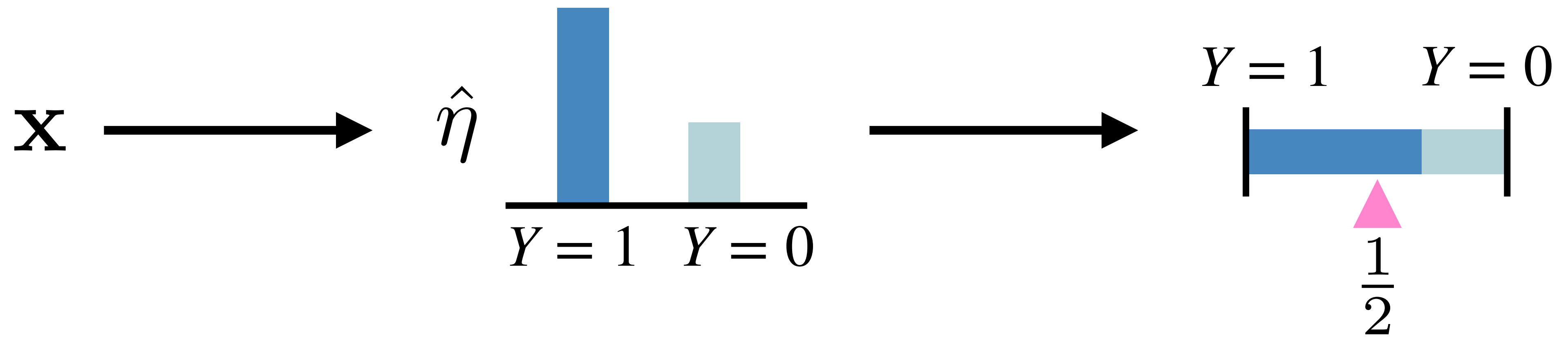
- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?



- How to derive surrogate regret bounds?

Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds

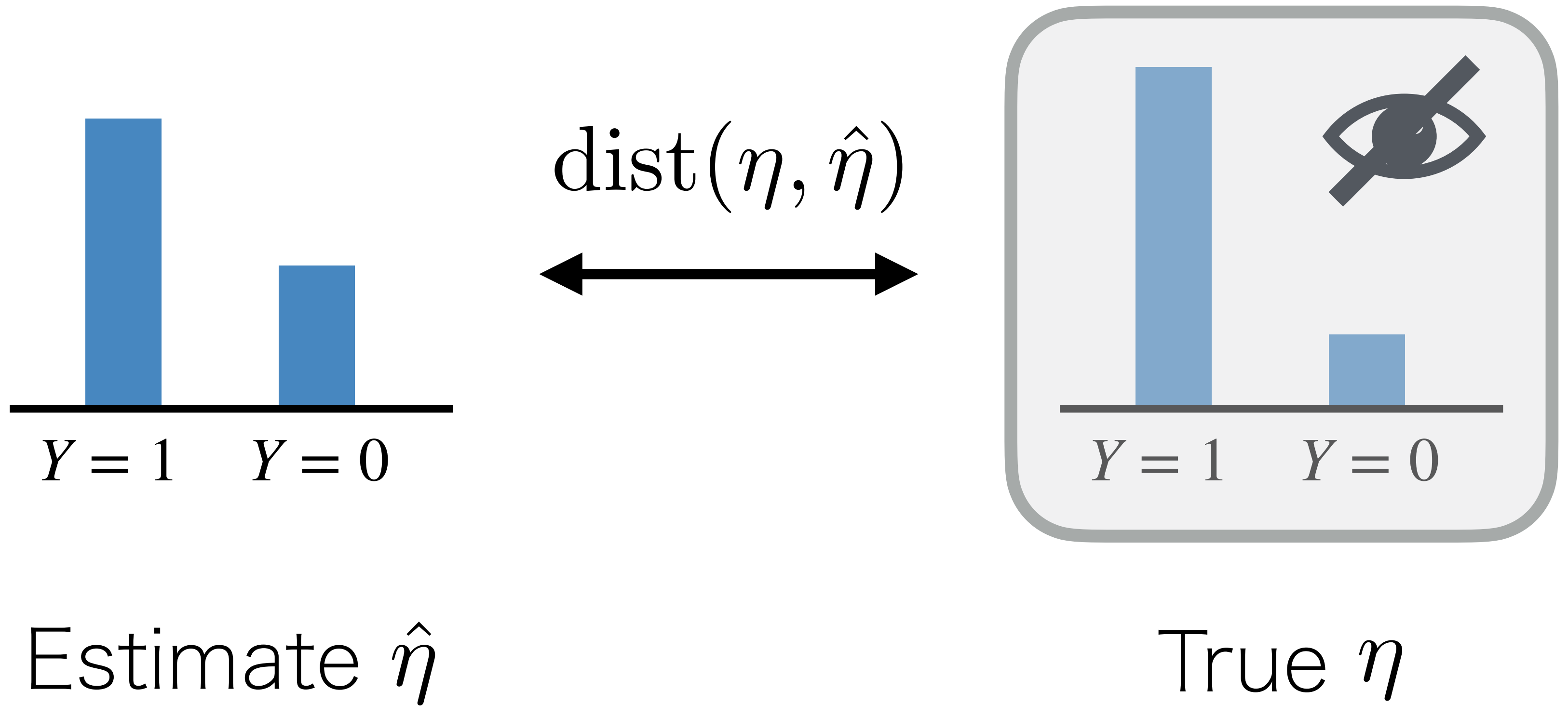
- How useful is probability estimate $\hat{\eta}$ for a downstream task?



Outline

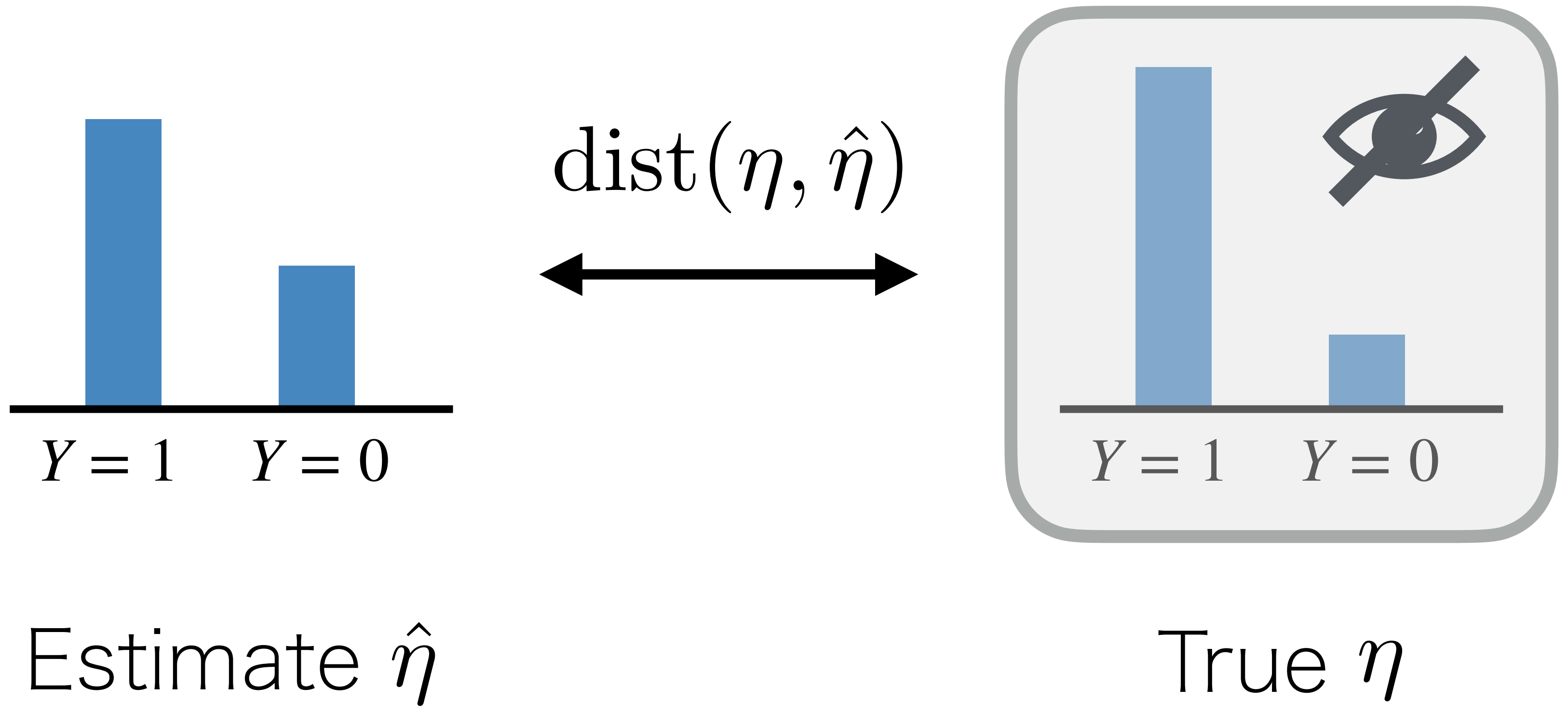
- **Q.** How should we assess probability estimates?
  - Proper losses
- **Q.** How can estimated probabilities be used for other tasks?
  - Regret bounds
- **Q.** How to compare different loss functions?
  - Order function of moduli

Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds



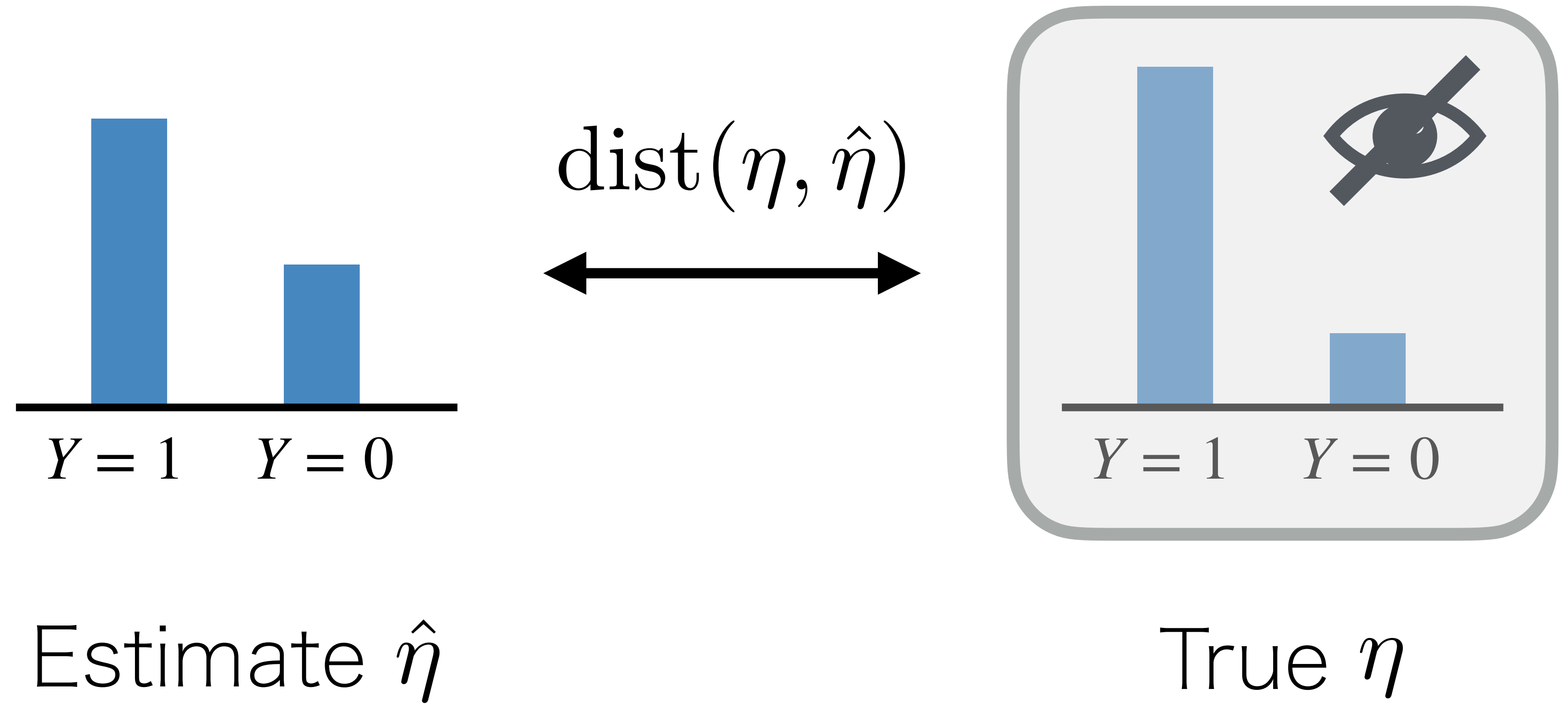
Proper losses

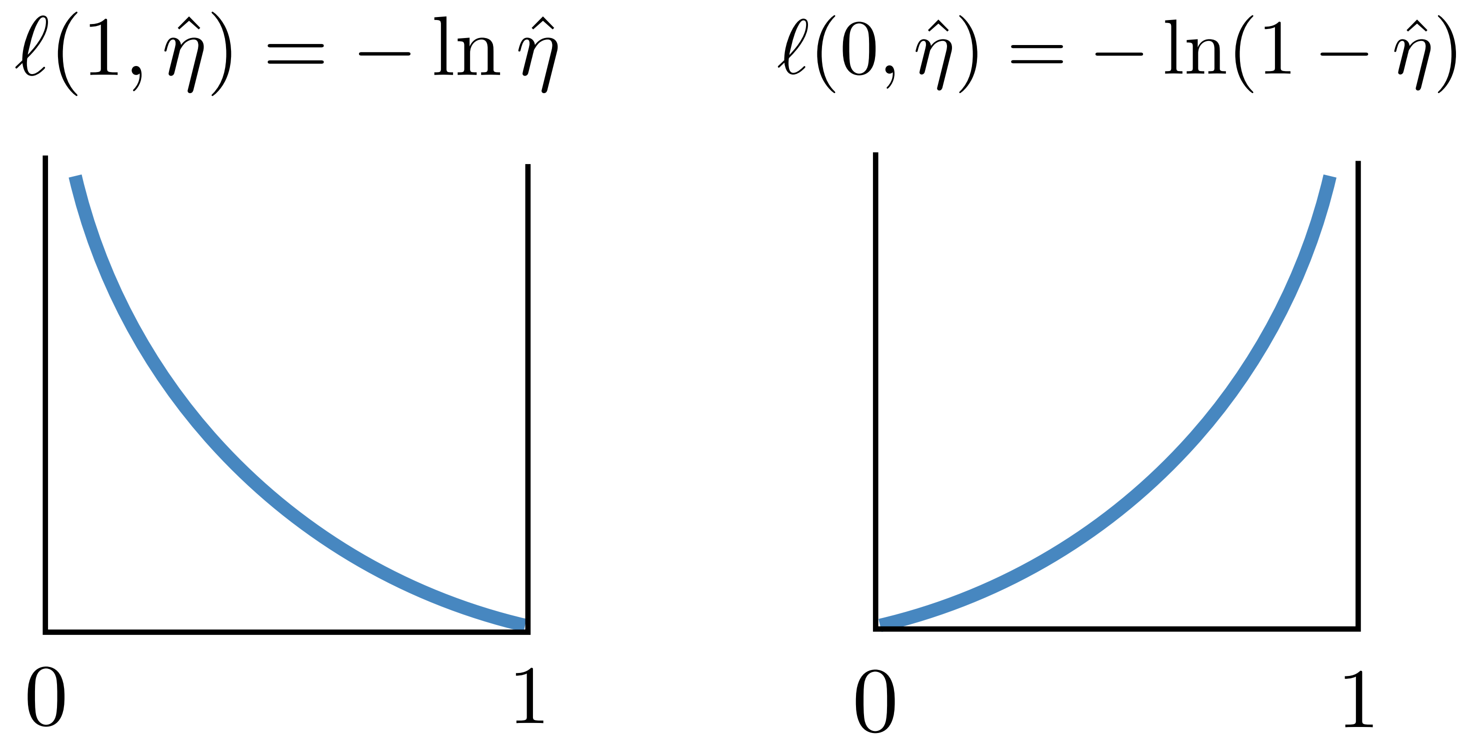
- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?
  - Challenge: no observation of $\eta(\mathbf{x}_i)$



Proper losses

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?
  - Challenge: no observation of $\eta(\mathbf{x}_i)$
- Loss function $\ell(y, \hat{\eta})$
  - Define loss function for $y \in \{1, 0\}$ separately
  - Example (log loss): $\ell(y, \hat{\eta}) = \begin{cases} -\ln \hat{\eta} & \text{if } y = 1 \\ -\ln(1 - \hat{\eta}) & \text{if } y = 0 \end{cases}$

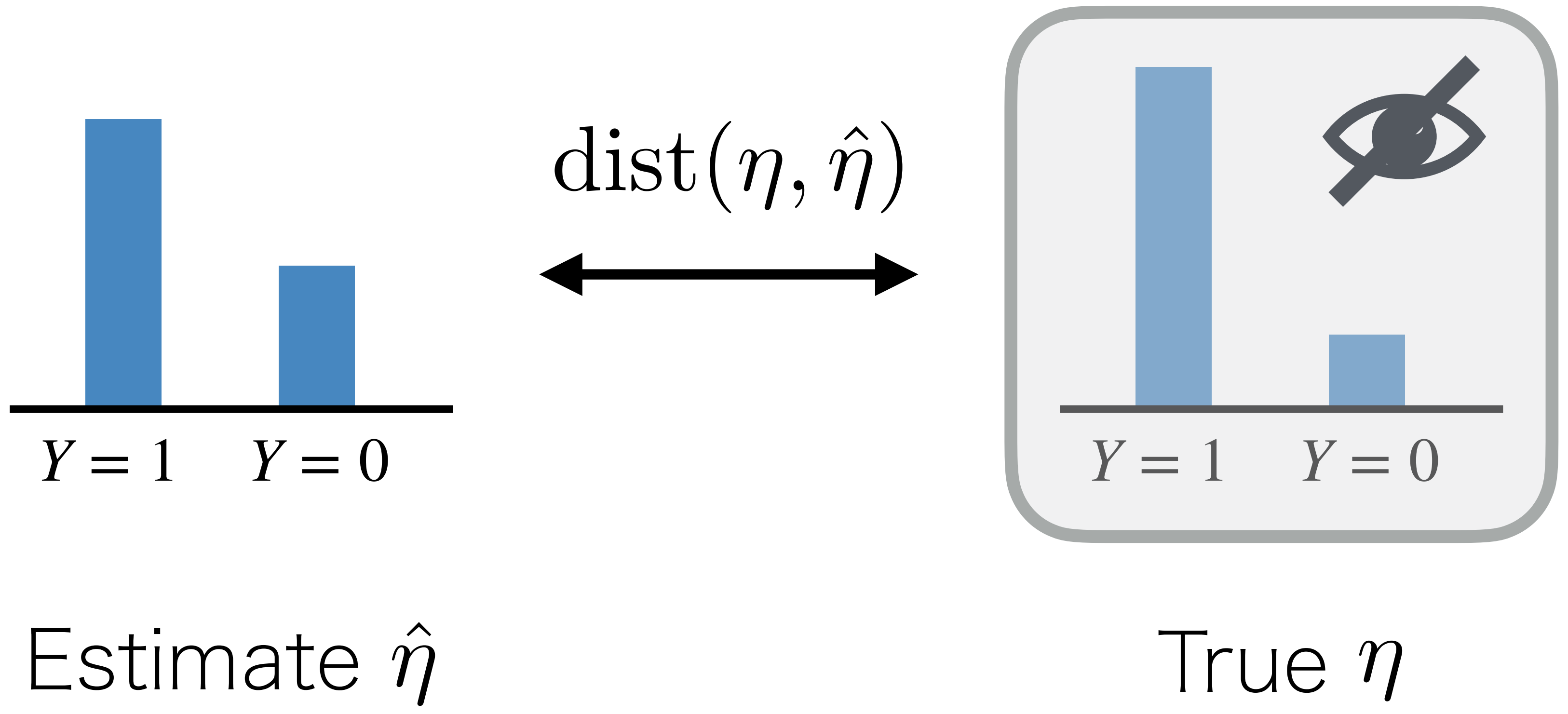


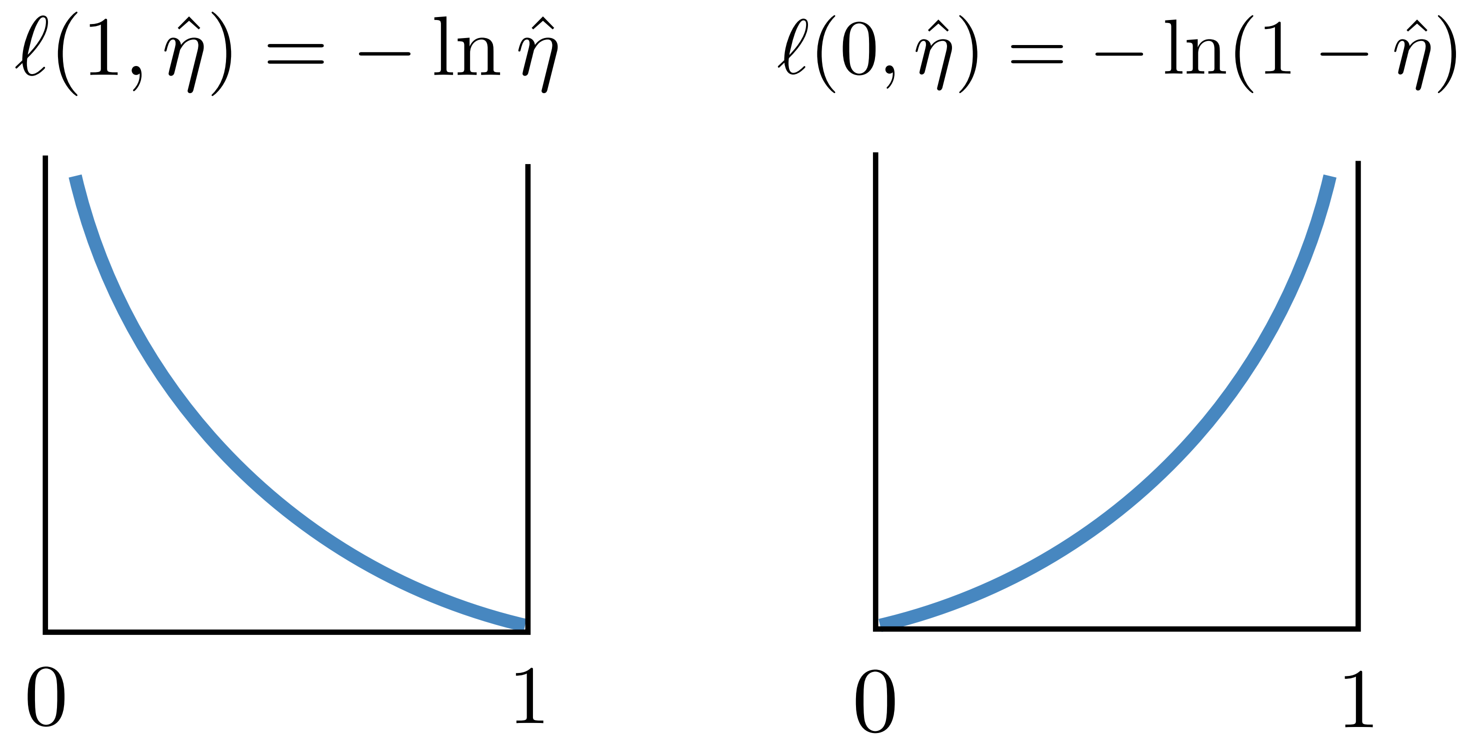


Proper losses

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?
  - Challenge: no observation of $\eta(\mathbf{x}_i)$
- Loss function $\ell(y, \hat{\eta})$
  - Define loss function for $y \in \{1, 0\}$ separately
  - Example (log loss): $\ell(y, \hat{\eta}) = \begin{cases} -\ln \hat{\eta} & \text{if } y = 1 \\ -\ln(1-\hat{\eta}) & \text{if } y = 0 \end{cases}$
- $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ can be assessed via expected loss

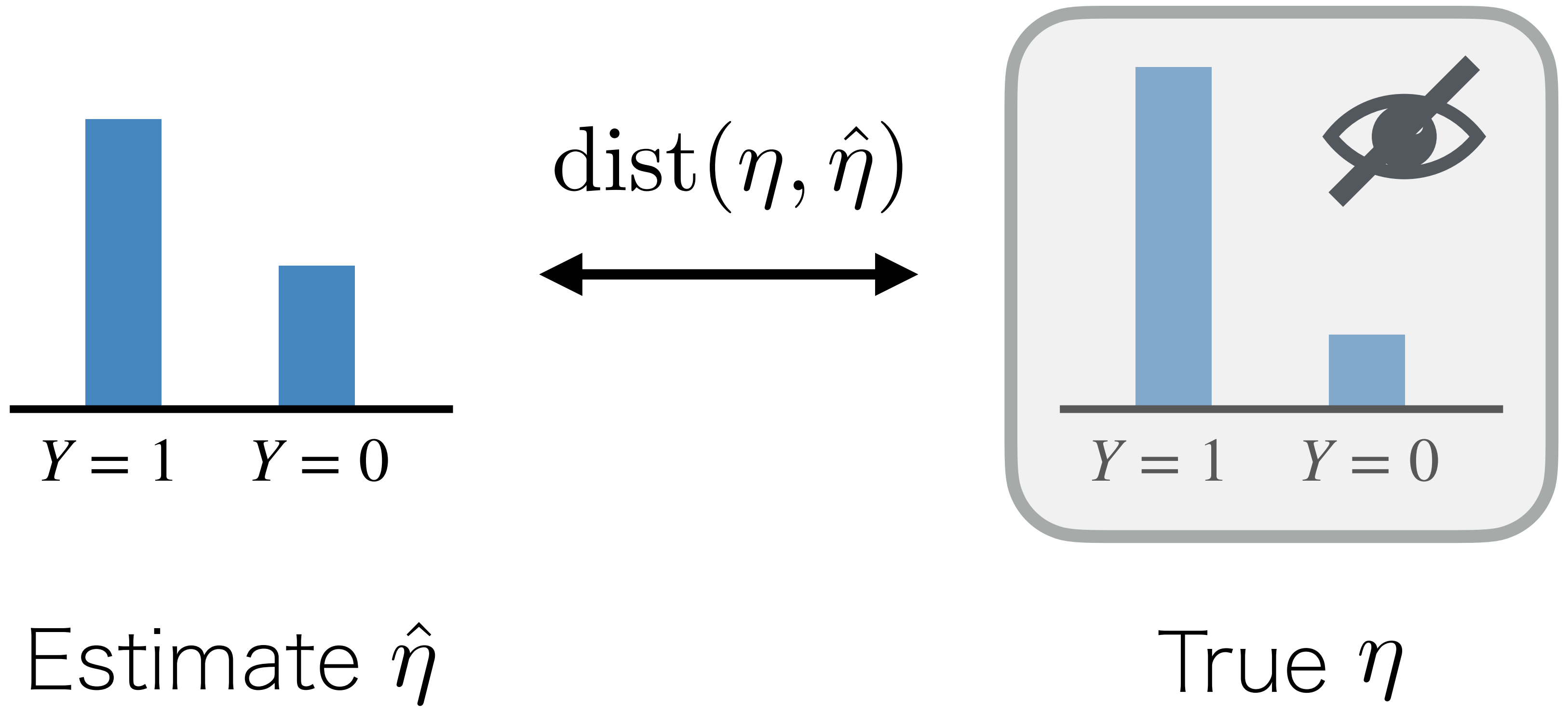
$$\mathbb{E}_{(X,Y)}\ell(Y, \hat{\eta}(X)) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\ell(Y, \hat{\eta}(X))\right]$$



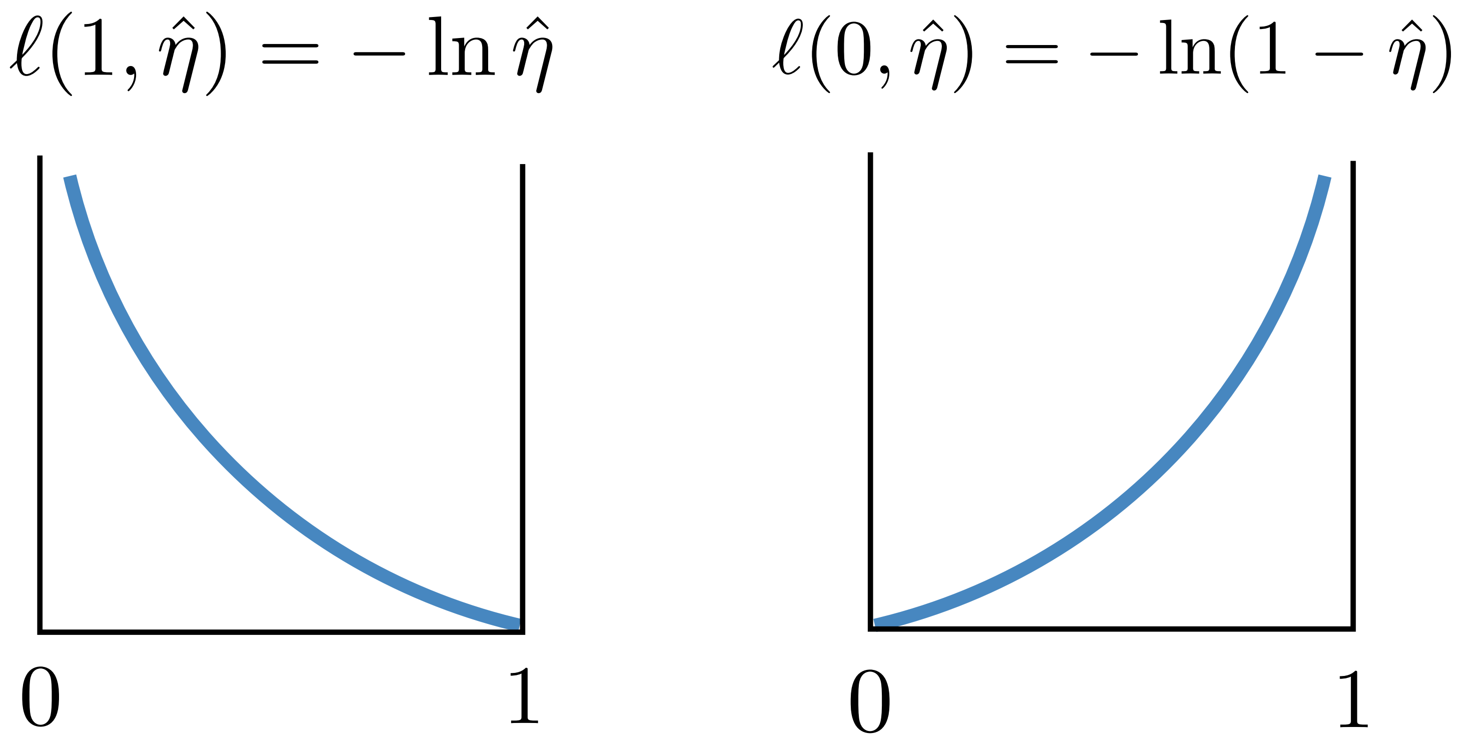


Proper losses

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?
  - Challenge: no observation of $\eta(\mathbf{x}_i)$



- Loss function $\ell(y, \hat{\eta})$
  - Define loss function for $y \in \{1, 0\}$ separately
  - Example (log loss): $\ell(y, \hat{\eta}) = \begin{cases} -\ln \hat{\eta} & \text{if } y = 1 \\ -\ln(1-\hat{\eta}) & \text{if } y = 0 \end{cases}$



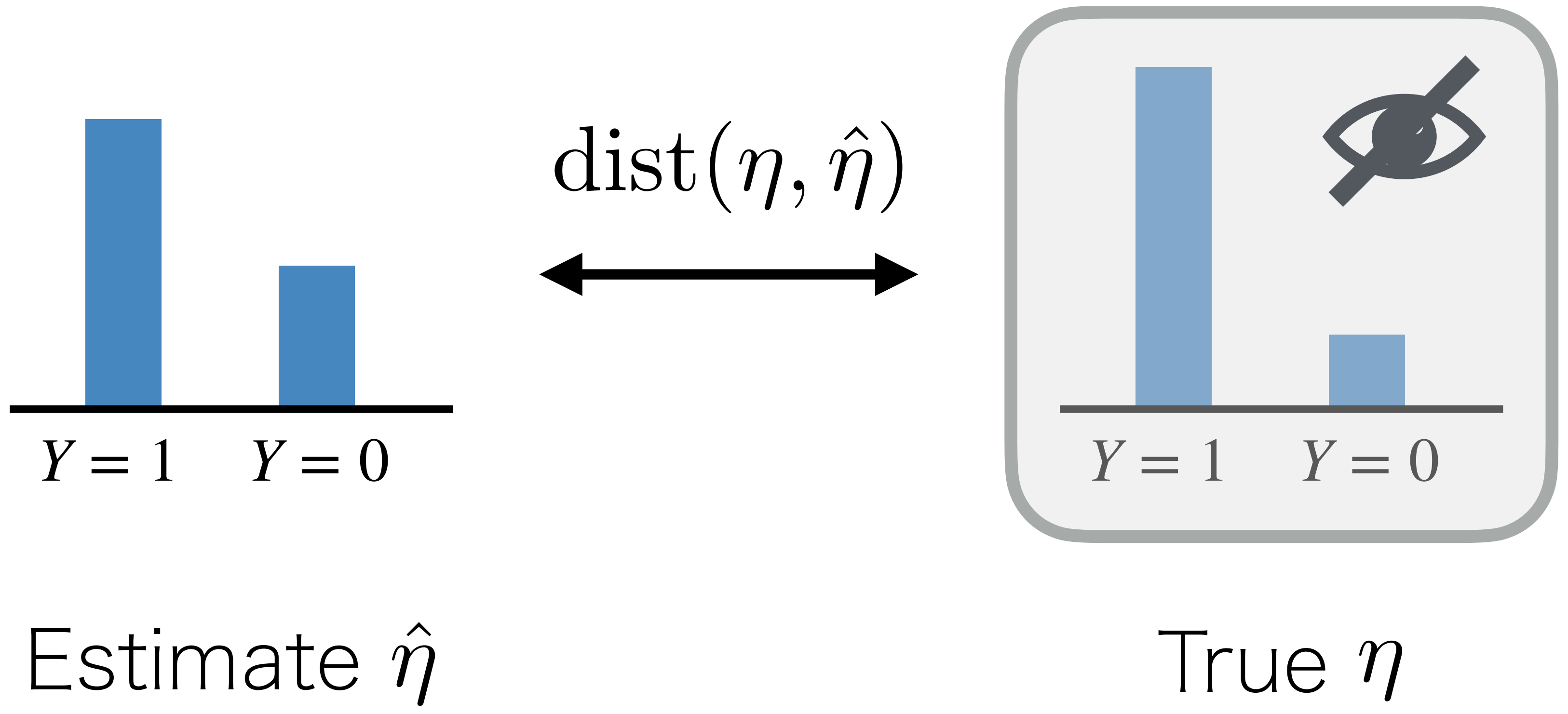
- $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ can be assessed via expected loss

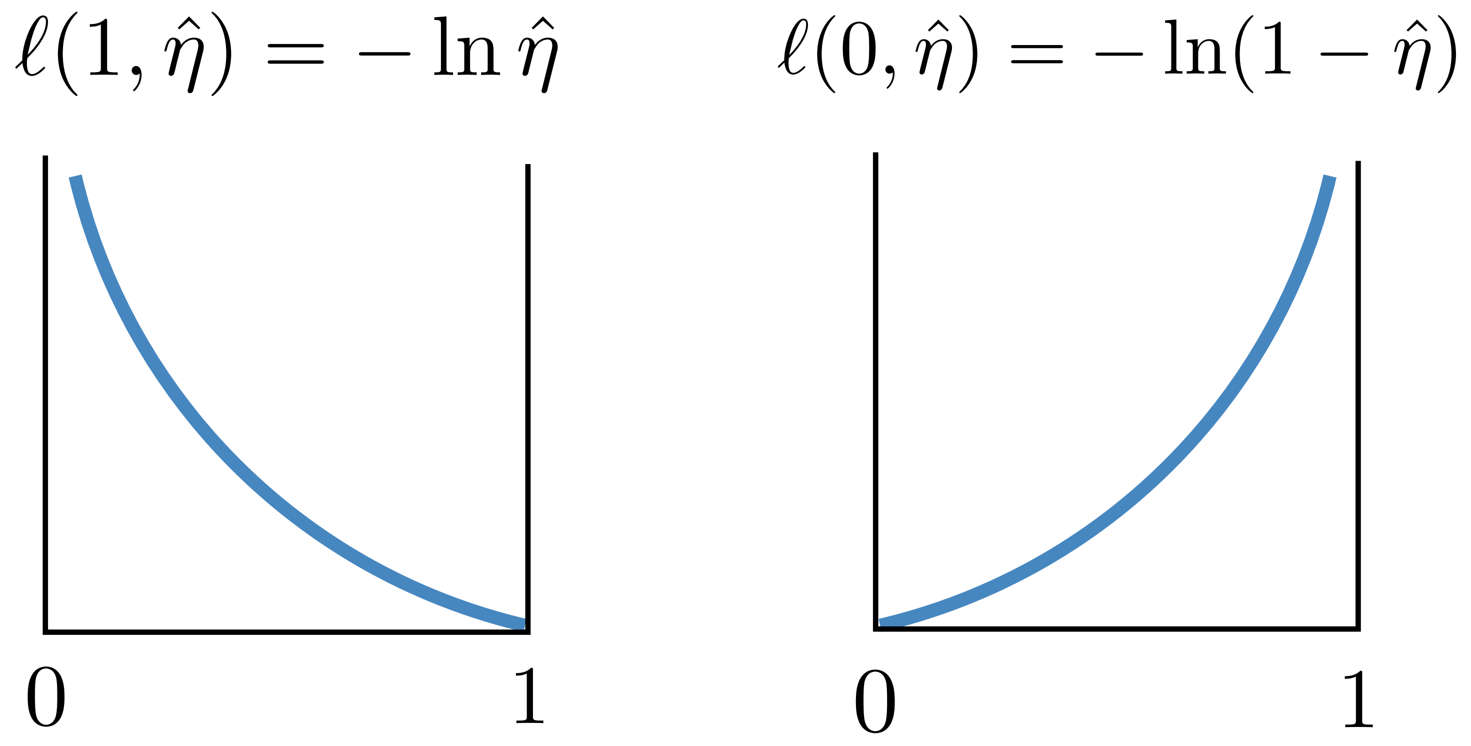
$$\mathbb{E}_{(X,Y)}\ell(Y, \hat{\eta}(X)) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\ell(Y, \hat{\eta}(X))\right]$$
$$= \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$$

Proper losses

- How to compute $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ with only $\hat{\eta}$ and y ?
  - Challenge: no observation of $\eta(\mathbf{x}_i)$
- Loss function $\ell(y, \hat{\eta})$
  - Define loss function for $y \in \{1, 0\}$ separately
  - Example (log loss): $\ell(y, \hat{\eta}) = \begin{cases} -\ln \hat{\eta} & \text{if } y = 1 \\ -\ln(1-\hat{\eta}) & \text{if } y = 0 \end{cases}$
- $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ can be assessed via expected loss

$$\mathbb{E}_{(X,Y)}\ell(Y, \hat{\eta}(X)) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\ell(Y, \hat{\eta}(X))\right]$$
$$= \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$$
$$= \mathrm{dist}(\eta, \hat{\eta})$$





Proper losses

- $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ can be assessed via expected loss $\mathbb{E}_{(X,Y)}\ell(Y, \hat{\eta}(X)) = \mathbb{E}_X\left[\mathbb{E}_{Y=1|X}\ell(Y, \hat{\eta}(X))\right]$

$$= \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$$

$$= \mathrm{dist}(\eta, \hat{\eta})$$

- **Q.** What conditions should $\ell(y, \hat{\eta})$ satisfy?
  - Minimization of $\mathrm{dist}(\eta, \hat{\eta})$ should lead to $\hat{\eta} \approx \eta$

Buja, Andreas, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. (2005).

Proper losses

- $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ can be assessed via expected loss $\mathbb{E}_{(X,Y)}\ell(Y, \hat{\eta}(X)) = \mathbb{E}_X\left[\mathbb{E}_{Y=1|X}\ell(Y, \hat{\eta}(X))\right]$

$$= \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$$
$$= \mathrm{dist}(\eta, \hat{\eta})$$

- **Q.** What conditions should $\ell(y, \hat{\eta})$ satisfy?
  - Minimization of $\mathrm{dist}(\eta, \hat{\eta})$ should lead to $\hat{\eta} \approx \eta$

Definition [Buja et al., 2005]**.** $\ell(y, \hat{\eta})$ is proper when $L_\ell(\eta, \eta) = \underline{L}_\ell(\eta)$ for all $\eta \in [0, 1]$.

Conditional risk $L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$

Bayes risk $\underline{L}_\ell(\eta) = \inf_{\hat{\eta} \in [0,1]} L_\ell(\eta, \hat{\eta})$

Buja, Andreas, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. (2005).

Proper losses

- $\mathrm{dist}(\eta(\mathbf{x}), \hat{\eta}(\mathbf{x}))$ can be assessed via expected loss $\mathbb{E}_{(X,Y)}\ell(Y, \hat{\eta}(X)) = \mathbb{E}_X\left[\mathbb{E}_{Y=1|X}\ell(Y, \hat{\eta}(X))\right]$

$$= \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$$
$$= \mathrm{dist}(\eta, \hat{\eta})$$

- **Q.** What conditions should $\ell(y, \hat{\eta})$ satisfy?
  - Minimization of $\mathrm{dist}(\eta, \hat{\eta})$ should lead to $\hat{\eta} \approx \eta$

Definition [Buja et al., 2005]**.** $\ell(y, \hat{\eta})$ is proper when $L_\ell(\eta, \eta) = \underline{L}_\ell(\eta)$ for all $\eta \in [0, 1]$.

Conditional risk $L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$

Bayes risk $\underline{L}_\ell(\eta) = \inf_{\hat{\eta}\in[0,1]} L_\ell(\eta, \hat{\eta})$

  - $\hat{\eta} = \eta$ minimizes conditional risk
  - We say $\ell(y, \hat{\eta})$ is strictly proper if the minimizer is unique

Buja, Andreas, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. (2005).

Proper losses

- **Q.** How to test properness?

Theorem [Savage 1971]. ℓ is proper iff $\underline{L}_\ell$ is concave and

$$L(\eta, \hat{\eta}) = \underline{L}_\ell(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'_\ell(\hat{\eta}) \text{ for all } \eta, \hat{\eta} \in (0, 1).$$

Theorem [Agarwal 2014]. ℓ is strictly proper iff $\underline{L}_\ell$ is strictly concave.

Definition. $\ell(y, \hat{\eta})$ is <u>strictly proper</u> iff $L_\ell(\eta, \hat{\eta}) = \underline{L}_\ell(\eta) \iff \hat{\eta} = \eta$ for all $\eta \in [0, 1]$.

$$L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1 - \eta)\ell(0, \hat{\eta}) \qquad \underline{L}_\ell(\eta) = \inf_{\hat{\eta} \in [0,1]} L_\ell(\eta, \hat{\eta})$$

Savage, Leonard J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66.336 (1971): 783-801.
Agarwal, Shivani. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research* 15 (2014): 1653-1674.

Proper losses

- **Q.** How to test properness?

Theorem [Savage 1971]. ℓ is proper iff $\underline{L}_\ell$ is concave and

$$L(\eta, \hat{\eta}) = \underline{L}_\ell(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'_\ell(\hat{\eta}) \text{ for all } \eta, \hat{\eta} \in (0, 1).$$

Theorem [Agarwal 2014]. ℓ is strictly proper iff $\underline{L}_\ell$ is strictly concave.

- [⇒] Check strict concavity of $\underline{L}_\ell$

Definition. $\ell(y, \hat{\eta})$ is strictly proper iff $L_\ell(\eta, \hat{\eta}) = \underline{L}_\ell(\eta) \iff \hat{\eta} = \eta$ for all $\eta \in [0, 1]$.

$$L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1 - \eta)\ell(0, \hat{\eta}) \qquad \underline{L}_\ell(\eta) = \inf_{\hat{\eta} \in [0,1]} L_\ell(\eta, \hat{\eta})$$

Savage, Leonard J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66.336 (1971): 783-801.
Agarwal, Shivani. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research* 15 (2014): 1653-1674.

Proper losses

- **Q.** How to test properness?

Theorem [Savage 1971]. ℓ is proper iff $\underline{L}_\ell$ is concave and

$$L(\eta, \hat{\eta}) = \underline{L}_\ell(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'_\ell(\hat{\eta}) \text{ for all } \eta, \hat{\eta} \in (0,1).$$

Theorem [Agarwal 2014]. ℓ is strictly proper iff $\underline{L}_\ell$ is strictly concave.

- [⇒] Check strict concavity of $\underline{L}_\ell$
- [⇐] For a concave $H : [0,1] \to \mathbb{R}$, loss $\ell(y, \hat{\eta}) = H(\hat{\eta}) + (y - \hat{\eta})H'(\hat{\eta})$ is proper
  - Remark: proper loss and concave function are closely related

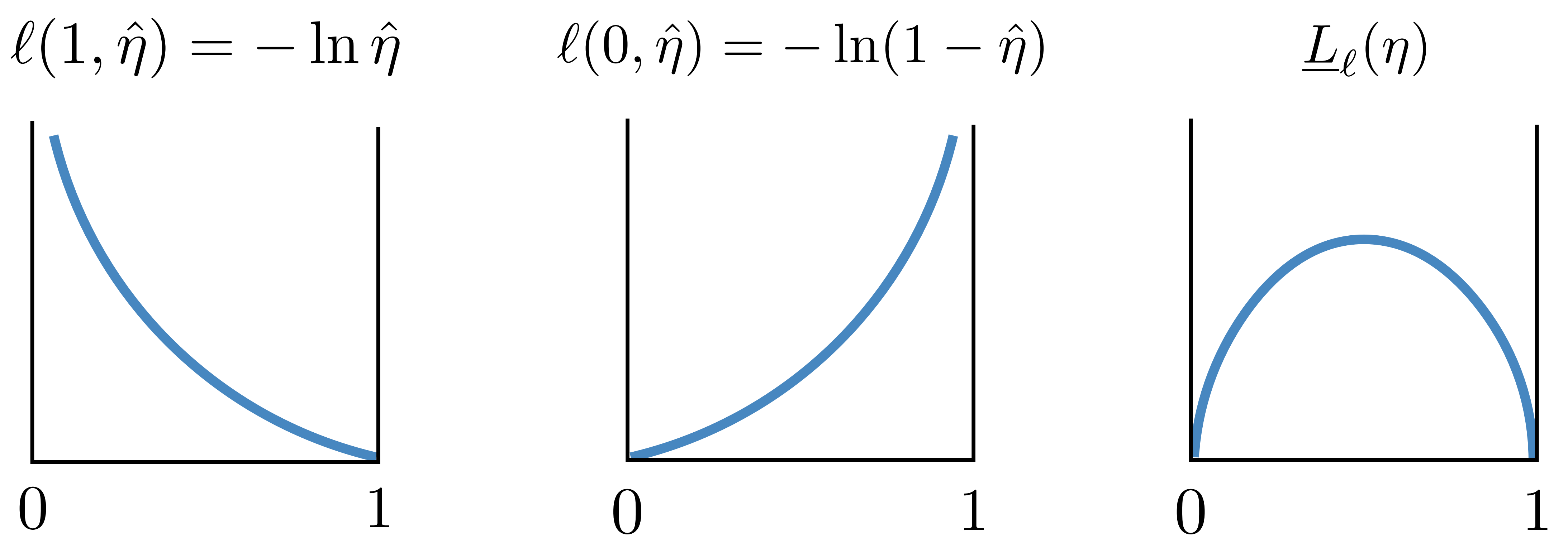
Definition. $\ell(y, \hat{\eta})$ is <u>strictly proper</u> iff $L_\ell(\eta, \hat{\eta}) = \underline{L}_\ell(\eta) \iff \hat{\eta} = \eta$ for all $\eta \in [0,1]$.

$$L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta}) \qquad \underline{L}_\ell(\eta) = \inf_{\hat{\eta} \in [0,1]} L_\ell(\eta, \hat{\eta})$$

Savage, Leonard J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66.336 (1971): 783-801.
Agarwal, Shivani. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research* 15 (2014): 1653-1674.

Examples | log loss

- Log loss $\ell(y,\hat\eta) = \begin{cases} -\ln\hat\eta & \text{if } y=1 \\ -\ln(1-\hat\eta) & \text{if } y=0 \end{cases}$
- **Conditional risk** $L_\ell(\eta,\hat\eta) = -\eta\ln\hat\eta - (1-\eta)\ln(1-\hat\eta)$
  - Binary cross-entropy
- **Bayes risk** $\underline{L}_\ell(\eta) = -\eta\ln\eta - (1-\eta)\ln(1-\eta)$
  - Shannon entropy
- Log loss is strictly proper because Shannon entropy is strictly concave

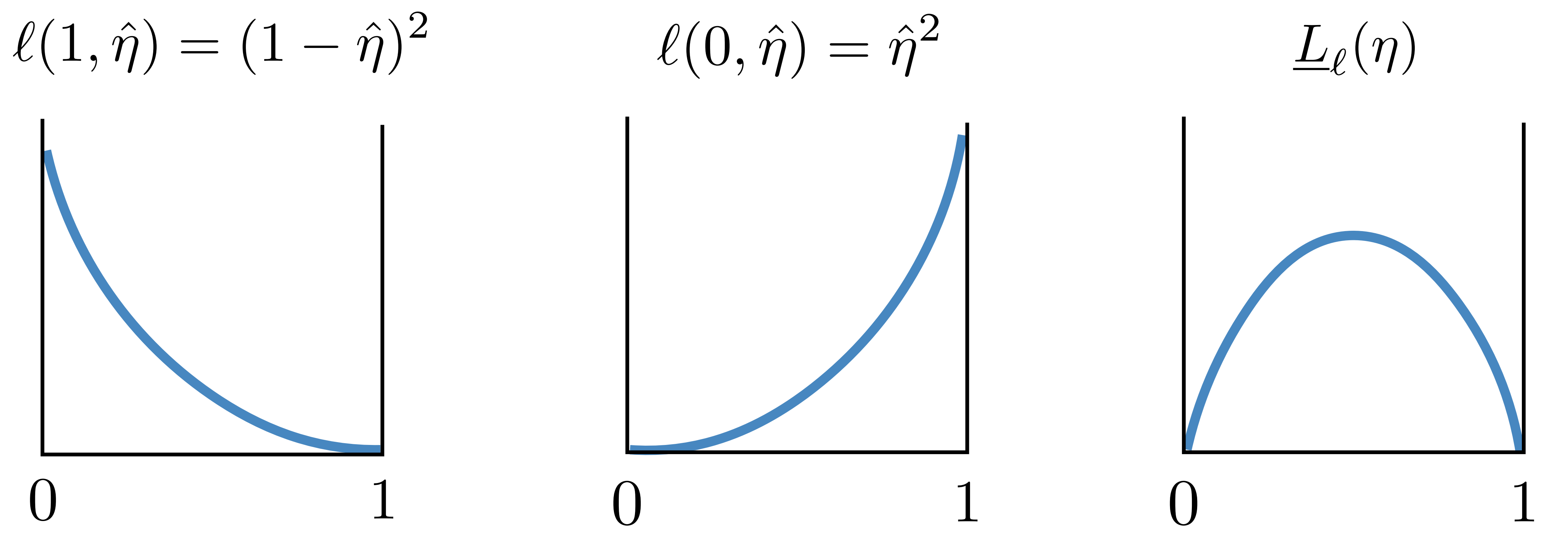


Definition. $\ell(y,\hat\eta)$ is <u>strictly proper</u> iff $L_\ell(\eta,\hat\eta) = \underline{L}_\ell(\eta) \iff \hat\eta = \eta$ for all $\eta \in [0,1]$.

$$L_\ell(\eta,\hat\eta) = \eta\ell(1,\hat\eta) + (1-\eta)\ell(0,\hat\eta) \qquad \underline{L}_\ell(\eta) = \inf_{\hat\eta\in[0,1]} L_\ell(\eta,\hat\eta)$$

Examples | L2 loss

- L2 loss $\ell(y, \hat{\eta}) = \begin{cases} (1-\hat{\eta})^2 & \text{if } y = 1 \\ \hat{\eta}^2 & \text{if } y = 0 \end{cases}$
- **Conditional risk** $L_\ell(\eta, \hat{\eta}) = \eta(1-\hat{\eta})^2 + (1-\eta)\hat{\eta}^2$

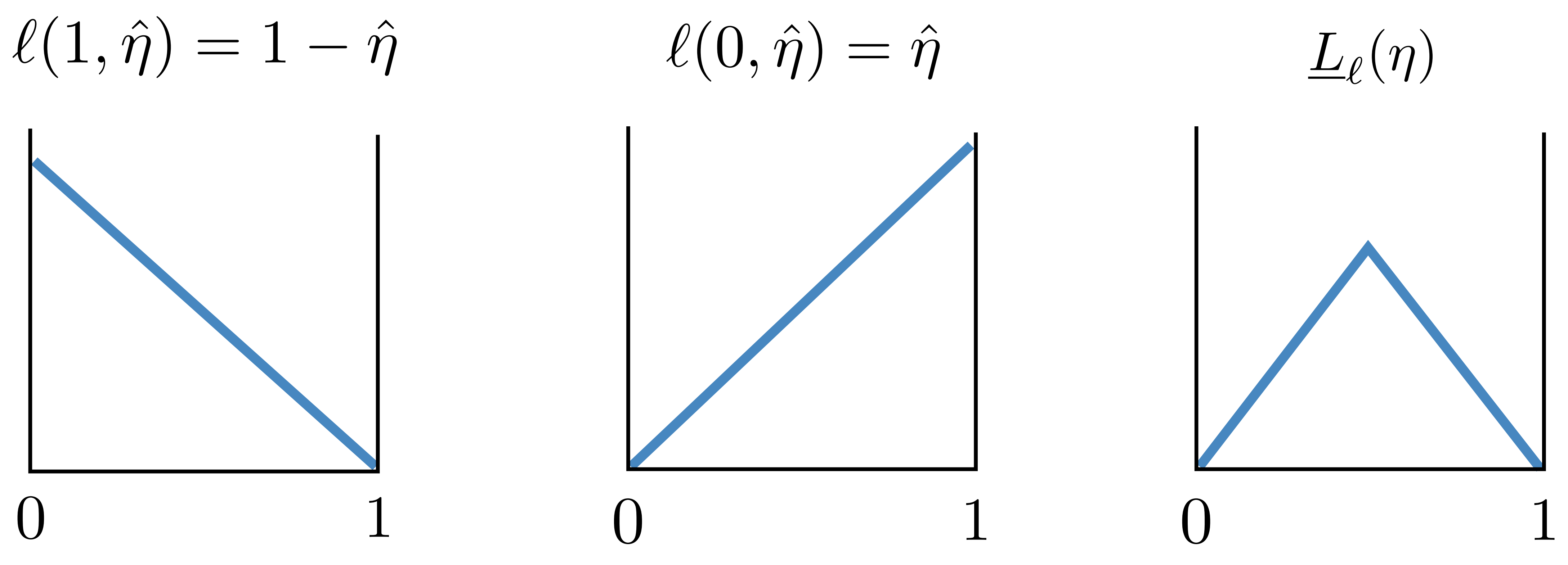


- **Bayes risk** $\underline{L}_\ell(\eta) = \eta(1-\eta)$
  - Gini index
- L2 loss is strictly proper because Gini index is strictly concave

Definition. $\ell(y, \hat{\eta})$ is strictly proper iff $L_\ell(\eta, \hat{\eta}) = \underline{L}_\ell(\eta) \iff \hat{\eta} = \eta$ for all $\eta \in [0, 1]$.

$$L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta}) \qquad \underline{L}_\ell(\eta) = \inf_{\hat{\eta}\in[0,1]} L_\ell(\eta, \hat{\eta})$$

Examples | L1 loss



- L1 loss $\ell(y,\hat\eta) = \begin{cases} 1-\hat\eta & \text{if } y=1 \\ \hat\eta & \text{if } y=0 \end{cases}$
- **Conditional risk** $L_\ell(\eta,\hat\eta) = \eta(1-\hat\eta) + (1-\eta)\hat\eta$
- **Bayes risk** $\underline{L}_\ell(\eta) = \max\{\eta, 1-\eta\}$
- L1 loss is not strictly proper **nor proper**

($\underline{L}_\ell(\eta)$ is not strictly concave)

$$L(\eta,\hat\eta) \neq \underline{L}_\ell(\hat\eta) + (\eta-\hat\eta)\underline{L}'_\ell(\hat\eta) = \begin{cases} \eta & \text{if } \hat\eta \in [0,\frac{1}{2}] \\ 1-\eta & \text{if } \hat\eta \in (\frac{1}{2},1] \end{cases}$$

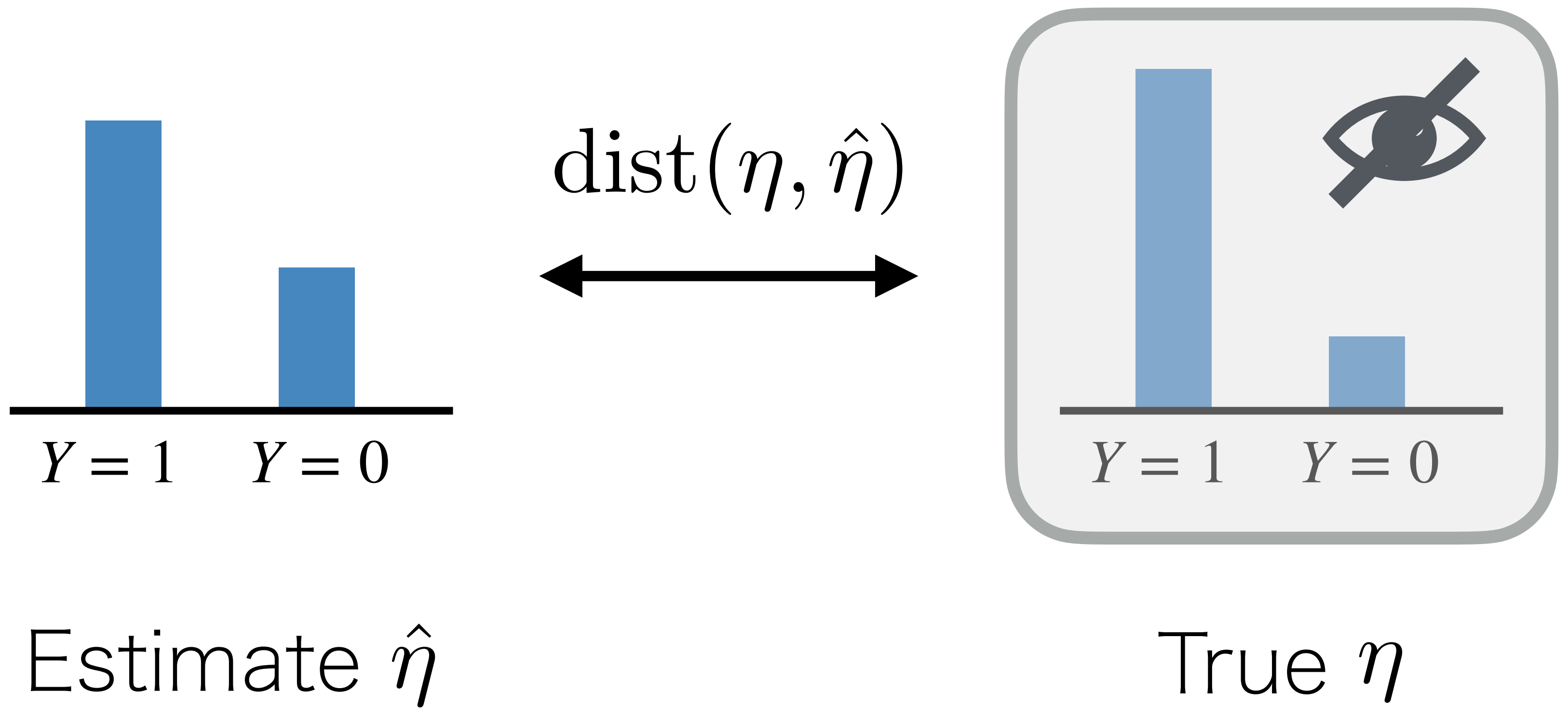
Theorem. ℓ is proper iff $\underline{L}_\ell$ is concave and $L(\eta,\hat\eta) = \underline{L}_\ell(\hat\eta) + (\eta-\hat\eta)\underline{L}'_\ell(\hat\eta)$.

Theorem. ℓ is strictly proper iff $\underline{L}_\ell$ is strictly concave.

Outline

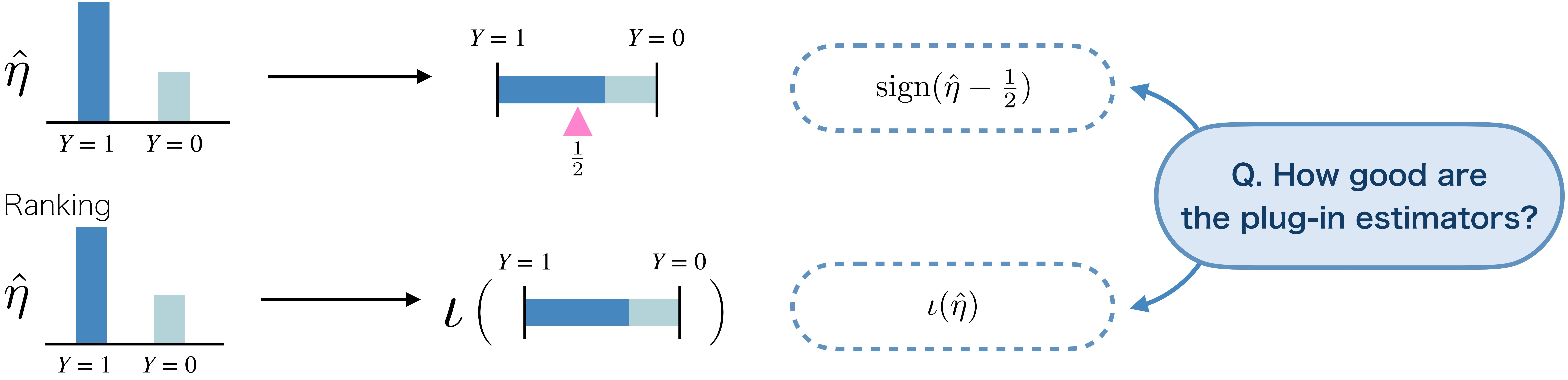
- **Q.** How should we assess probability estimates?
  - Proper losses
- **Q.** How can estimated probabilities be used for other tasks?
  - Regret bounds
- **Q.** How to compare different loss functions?
  - Order function of moduli

Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds



Proper loss vs. downstream tasks?

- Formulation of probability estimation: $\min_{\hat{\eta}:\mathcal{X}\to[0,1]} \mathbb{E}_X\left[L_\ell(\eta(X), \hat{\eta}(X))\right]$
  - We focus on "pointwise" problem $\min_{\hat{\eta}\in[0,1]} L_\ell(\eta, \hat{\eta})$
  - Equivalent to minimizing **regret** $R_\ell(\eta, \hat{\eta}) = L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta)$
- **Q.** How much does estimated $\hat{\eta}$ perform well for downstream tasks?
  - Classification
  - Ranking



How to relate a proper loss with downstream tasks?

Out attempt: to derive **L1 regret bound**

$$|\eta - \hat{\eta}| \leq \psi(R_\ell(\eta, \hat{\eta})) \text{ for regret } R_\ell(\eta, \hat{\eta}) = L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta)$$

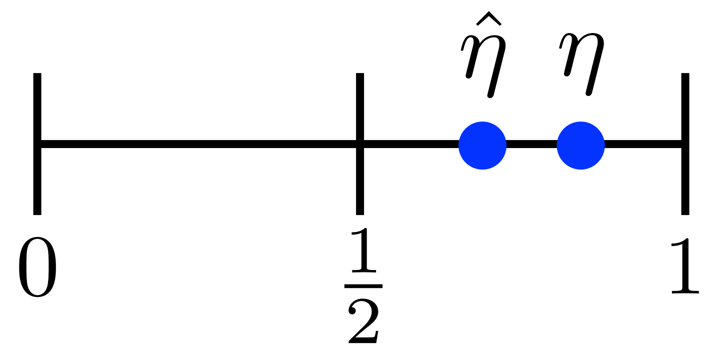
- Why? Because the optimality of downstream tasks can be bounded by $|\eta - \hat{\eta}|$

How to relate a proper loss with downstream tasks?

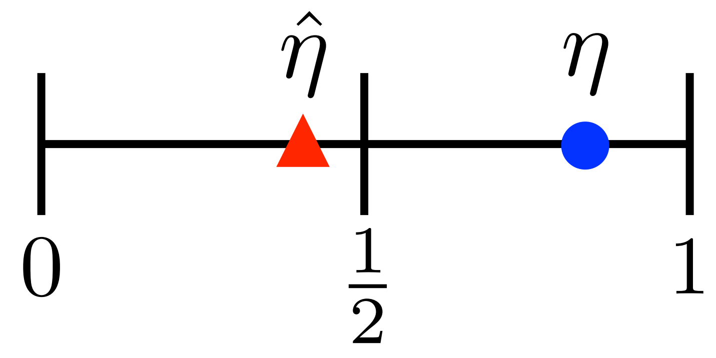
Out attempt: to derive **L1 regret bound**

$$|\eta - \hat{\eta}| \le \psi(R_\ell(\eta, \hat{\eta})) \text{ for regret } R_\ell(\eta, \hat{\eta}) = L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta)$$

- Why? Because the optimality of downstream tasks can be bounded by $|\eta - \hat{\eta}|$
- Classification $R_{01}(\eta, \hat{\eta}) = \left|\eta - \frac{1}{2}\right| [\![\min\{\eta, \hat{\eta}\} \le \frac{1}{2} < \max\{\eta, \hat{\eta}\}]\!] \le |\eta - \hat{\eta}|$ [Menon et al., 2013]



Not penalized



Penalized

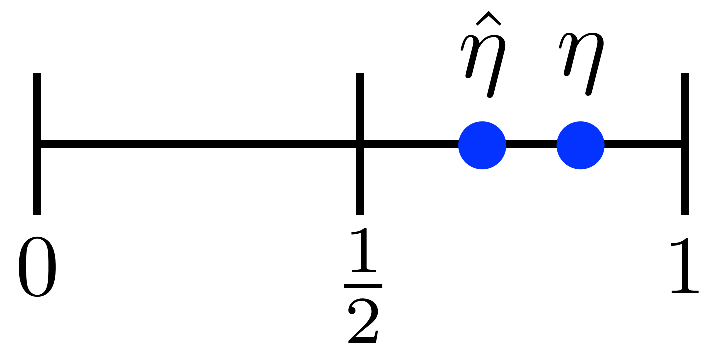
Menon, Aditya Krishna, et al. On the statistical consistency of algorithms for binary classification under class imbalance. *ICML* (2013).

How to relate a proper loss with downstream tasks?

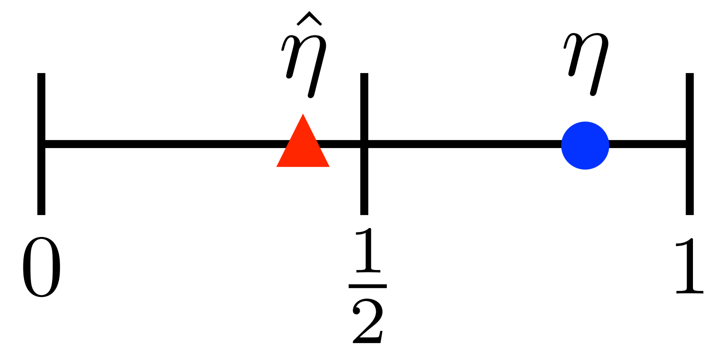
Out attempt: to derive **L1 regret bound**

$$|\eta - \hat{\eta}| \le \psi(R_\ell(\eta, \hat{\eta})) \text{ for regret } R_\ell(\eta, \hat{\eta}) = L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta)$$

- Why? Because the optimality of downstream tasks can be bounded by $|\eta - \hat{\eta}|$
- Classification $R_{01}(\eta, \hat{\eta}) = |\eta - \frac{1}{2}| [\![\min\{\eta, \hat{\eta}\} \le \frac{1}{2} < \max\{\eta, \hat{\eta}\}]\!] \le |\eta - \hat{\eta}|$ [Menon et al., 2013]

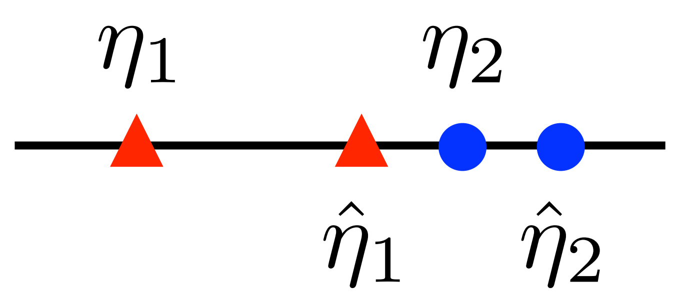


Not penalized

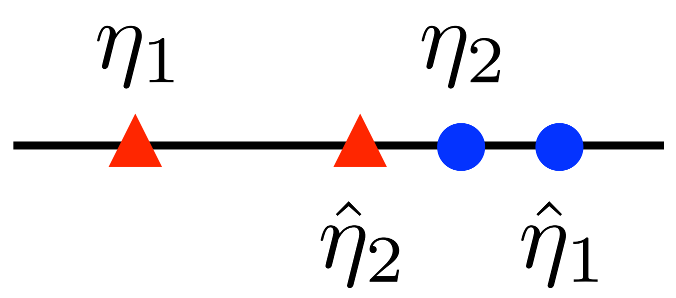


Penalized

- Ranking $R_{\text{rank}}(\eta_1, \eta_2, \hat{\eta}_1, \hat{\eta}_2) = |\eta_1 - \eta_2| [\![(\hat{\eta}_1 - \hat{\eta}_2)(\eta_1 - \eta_2) < 0]\!] \le |\eta_1 - \hat{\eta}_1| + |\eta_2 - \hat{\eta}_2|$ [Agarwal 2014]



Not penalized



Penalized

Menon, Aditya Krishna, et al. On the statistical consistency of algorithms for binary classification under class imbalance. *ICML* (2013).
Agarwal, Shivani. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research* 15 (2014): 1653-1674.

Motivation of our work

Out attempt: to derive **L1 regret bound**

$$|\eta - \hat{\eta}| \leq \psi(R_\ell(\eta, \hat{\eta})) \text{ for regret } R_\ell(\eta, \hat{\eta}) = L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta)$$

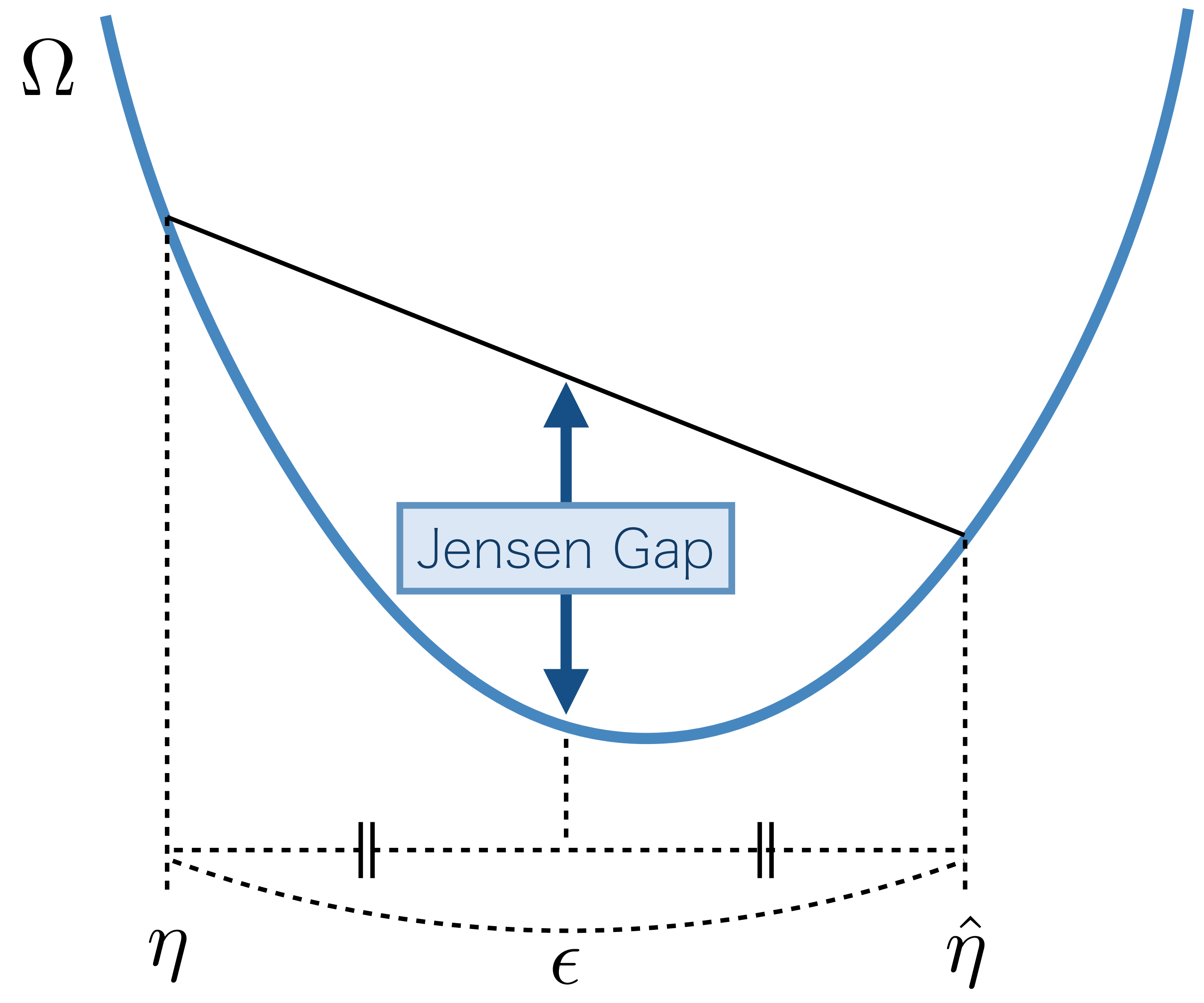
- Regret bound reads:
  minimizing regret R_ℓ amounts to be optimizing downstream performance via $|\eta - \hat{\eta}|$
- Motivation 1: to avoid deriving regret bounds for each task independently
- Motivation 2: to get more insight into relationship between a proper loss and ψ
- Consequently, we ask
  - Any loss performs universally well for various tasks?
  - Which loss entails faster rate?

Preparation | Moduli of convexity

Definition. For a convex function $\Omega : [0,1] \to \mathbb{R}$, its modulus of convexity is

$$\delta_\Omega(\epsilon) := \inf_{\eta, \hat{\eta} \in [0,1]} \left\{ \frac{\Omega(\eta) + \Omega(\hat{\eta})}{2} - \Omega\left(\frac{\eta + \hat{\eta}}{2}\right) \middle| \, |\eta - \hat{\eta}| \geq \epsilon \right\}.$$

- Moduli = the minimum Jensen gap of a convex function
  - Ω is strictly convex iff $\delta_\Omega(\epsilon) > 0 \quad \forall \epsilon > 0$
  - Generalization of strongly convex functions with $\delta_\Omega(\epsilon) = \frac{\mu}{2}\epsilon^2$



Main theorem | Regret upper bounds

Theorem. For a proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, for all $\eta, \hat{\eta} \in [0,1]$,

$$\delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|) \leq R_\ell(\eta, \hat{\eta}).$$

- Monotone function $\delta_{-\underline{L}_\ell}$ governs L1 regret bound ($-\underline{L}_\ell$: convex)
- **Corollary.** Regret bounds for downstream tasks
  - Classification $R_{01}(\eta, \hat{\eta}) \leq (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}(R_\ell(\eta, \hat{\eta}))$
  - Ranking $R_{\text{rank}}(\eta_1, \eta_2, \hat{\eta}_1, \hat{\eta}_2) \leq (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}(R_\ell(\eta_1, \hat{\eta}_1)) + (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}(R_\ell(\eta_2, \hat{\eta}_2))$

($\delta^{\star\star}_\Omega$ is convex biconjugate, and hence convex)

Regret $R_\ell(\eta, \hat{\eta}) = L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta)$

Conditional risk $L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1-\eta)\ell(0, \hat{\eta})$

Bayes risk $\underline{L}_\ell(\eta) = \inf_{\hat{\eta} \in [0,1]} L_\ell(\eta, \hat{\eta})$

Proof sketch

Theorem. For a proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, for all $\eta, \hat{\eta} \in [0,1]$,

$$\delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|) \leq R_\ell(\eta, \hat{\eta}).$$

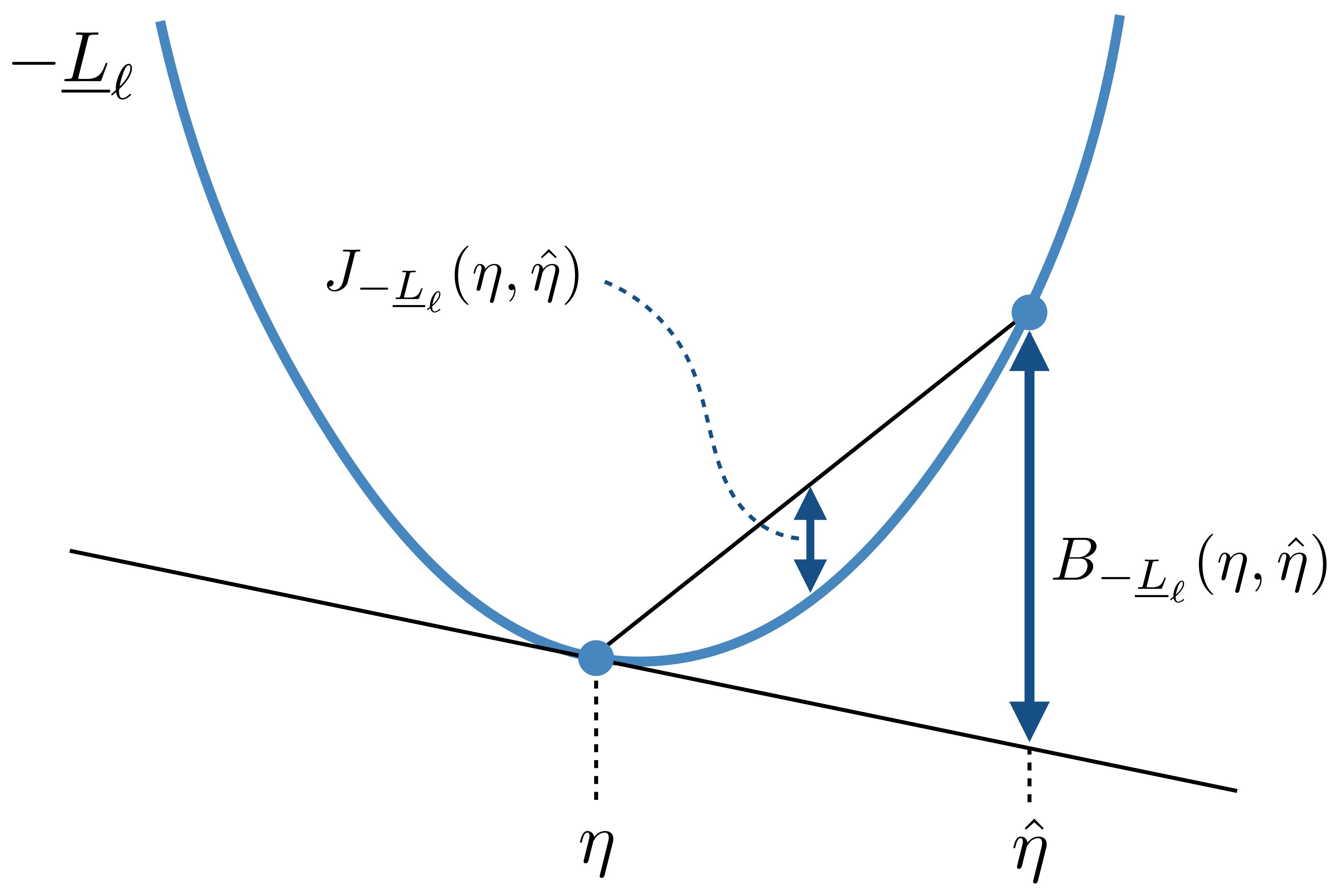
- Modulus is bounded by **Jensen-Bregman divergence**

$$\delta_{-\underline{L}_\ell}(\epsilon) \leq \frac{-\underline{L}_\ell(\eta) - \underline{L}_\ell(\hat{\eta})}{2} + \underline{L}_\ell\left(\frac{\eta + \hat{\eta}}{2}\right) =: J_{-\underline{L}_\ell}(\eta, \hat{\eta})$$

- Regret of proper loss = **Bregman divergence**

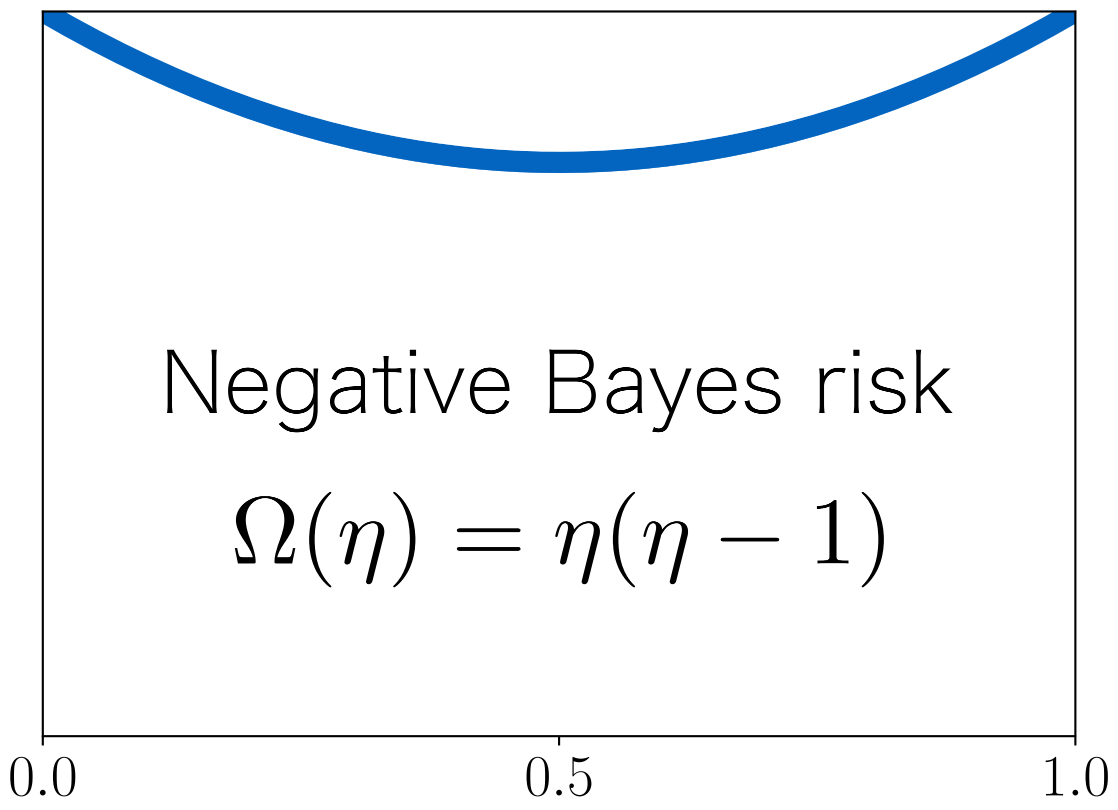
$$R_\ell(\eta, \hat{\eta}) = -\underline{L}_\ell(\hat{\eta}) + \underline{L}_\ell(\eta) + (\hat{\eta} - \eta)\underline{L}_\ell'(\eta) =: B_{-\underline{L}_\ell}(\eta, \hat{\eta})$$

- It is sufficient to show $J_{-\underline{L}_\ell}(\eta, \hat{\eta}) \leq B_{-\underline{L}_\ell}(\eta, \hat{\eta})$
  - ❖ Evident from figure



Examples

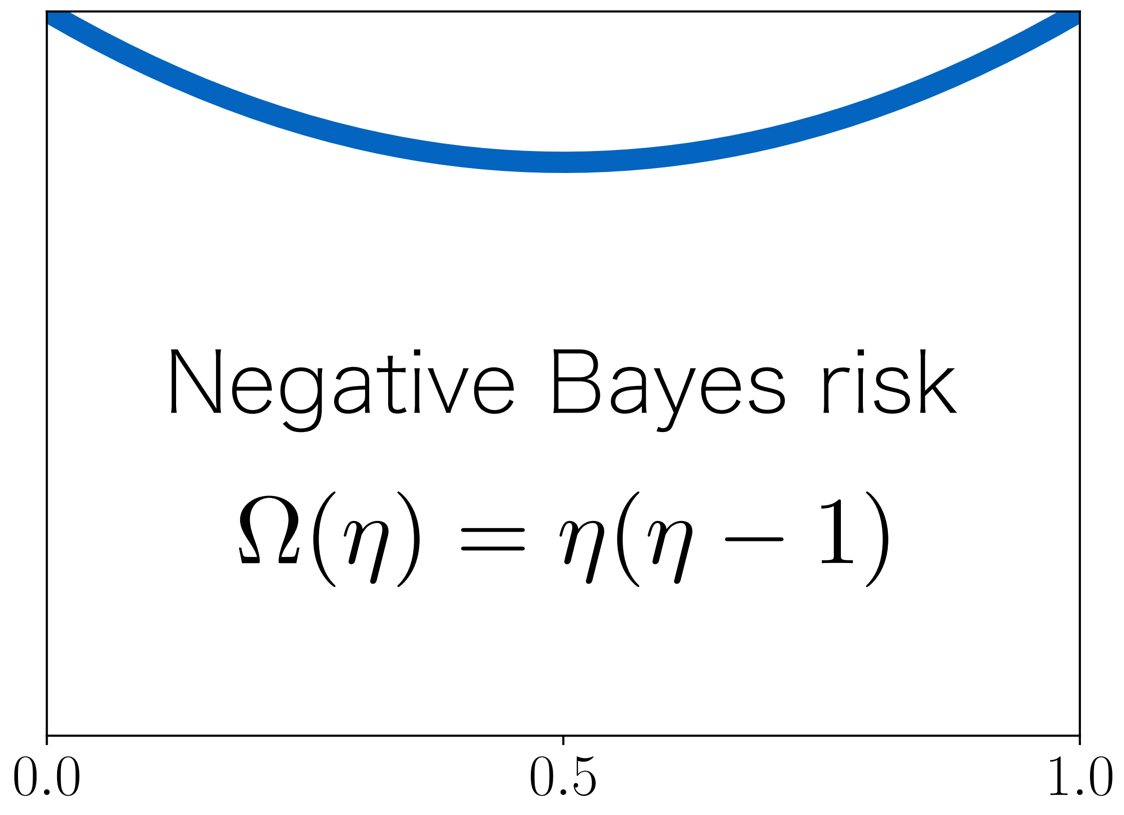
- L2 loss

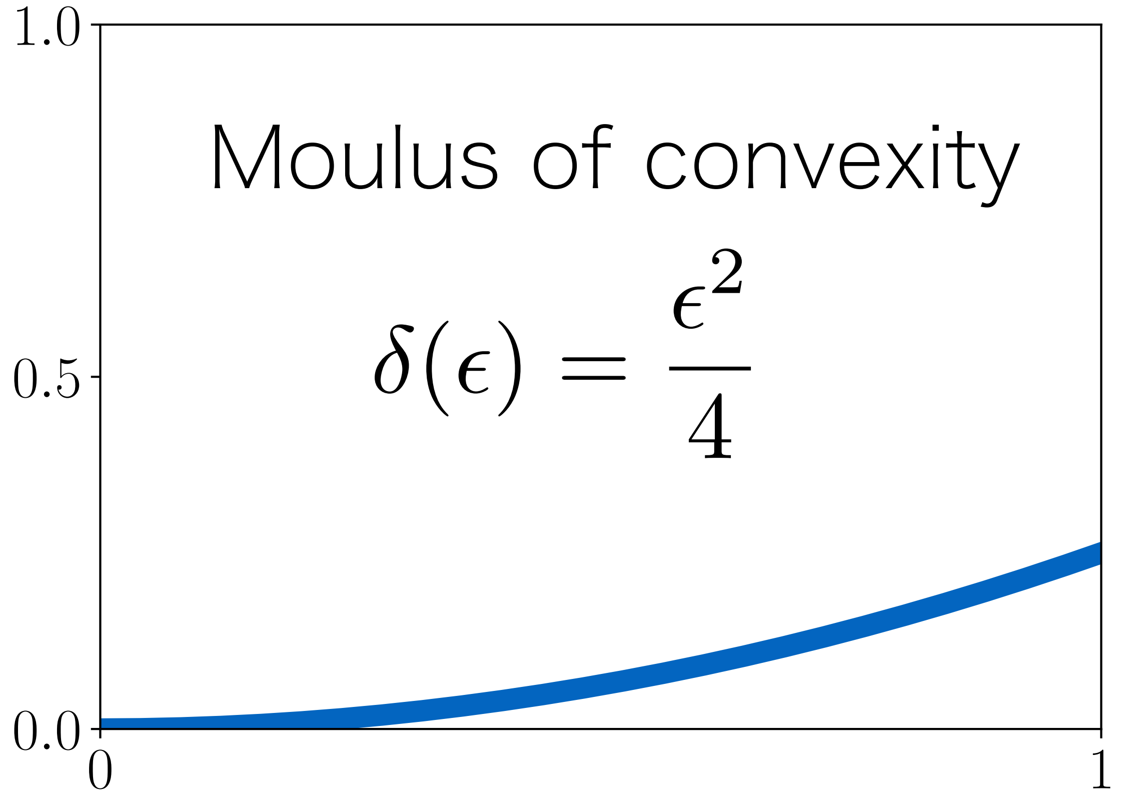


$(\ \Omega = -\underline{L}_\ell\ )$

Examples

- L2 loss

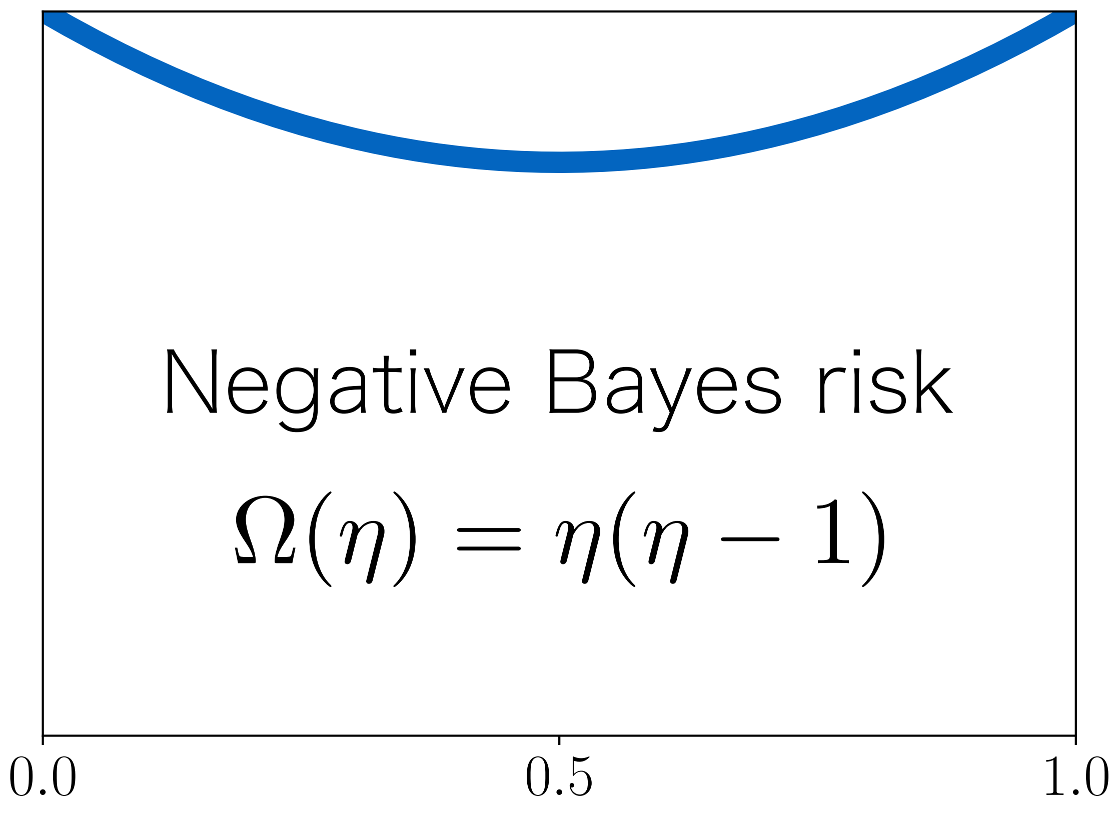




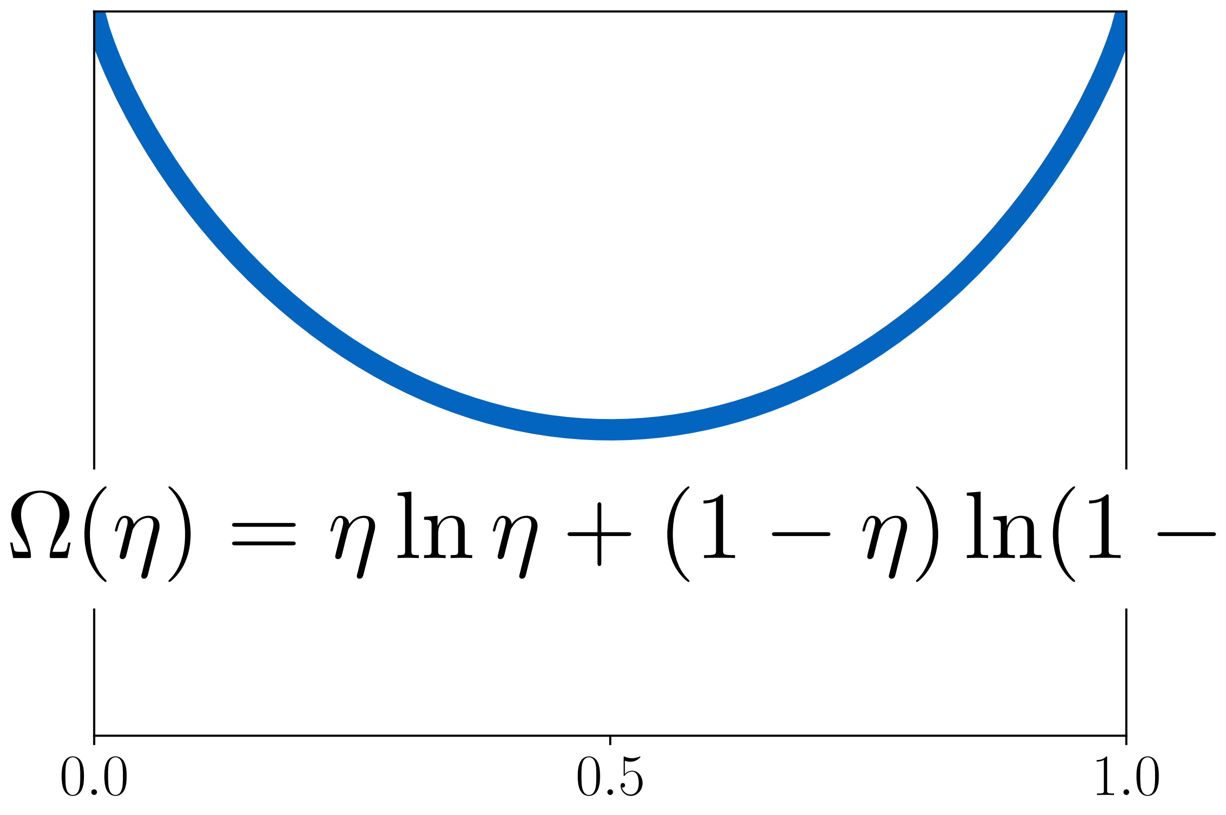
($\Omega = -\underline{L}_\ell$)

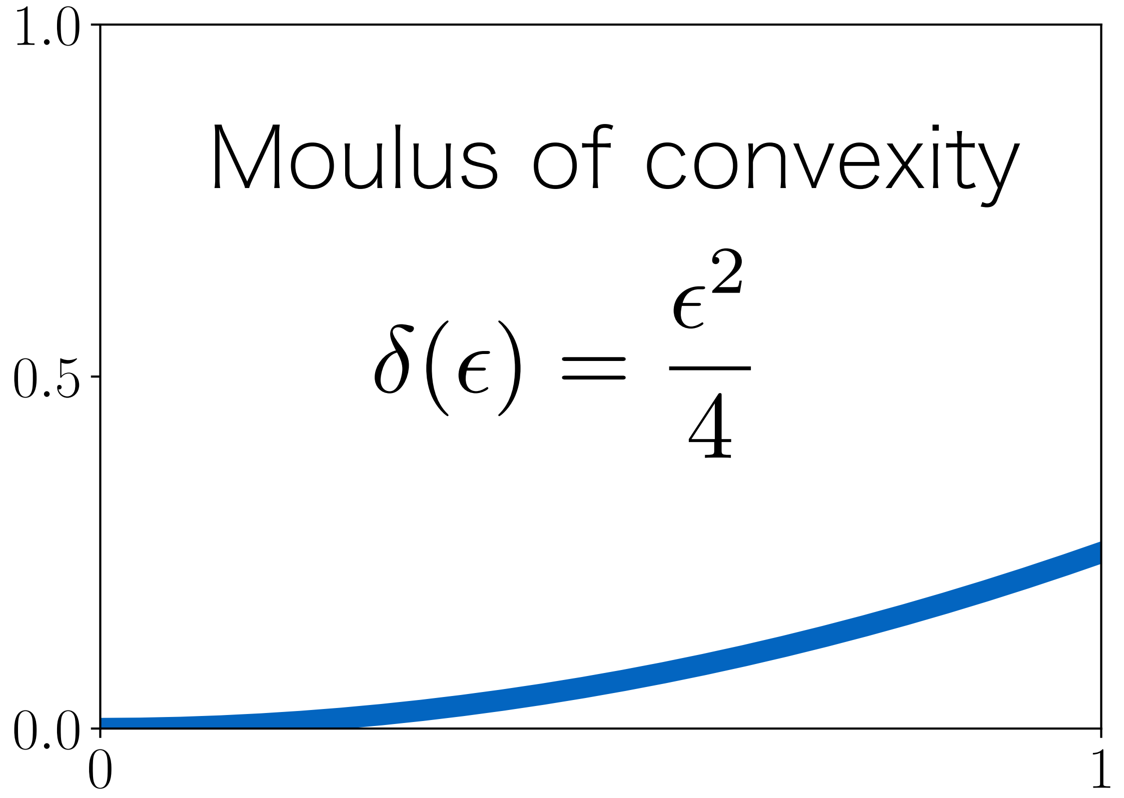
Examples

- L2 loss



- Log loss

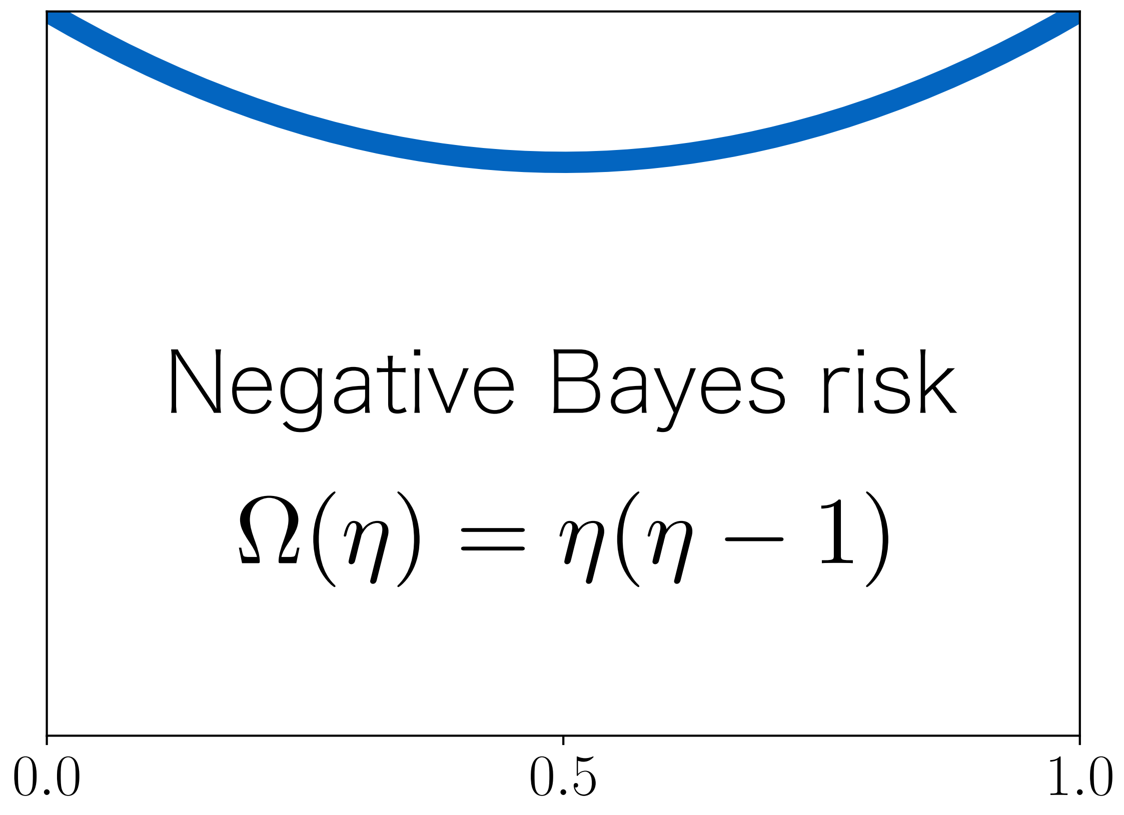


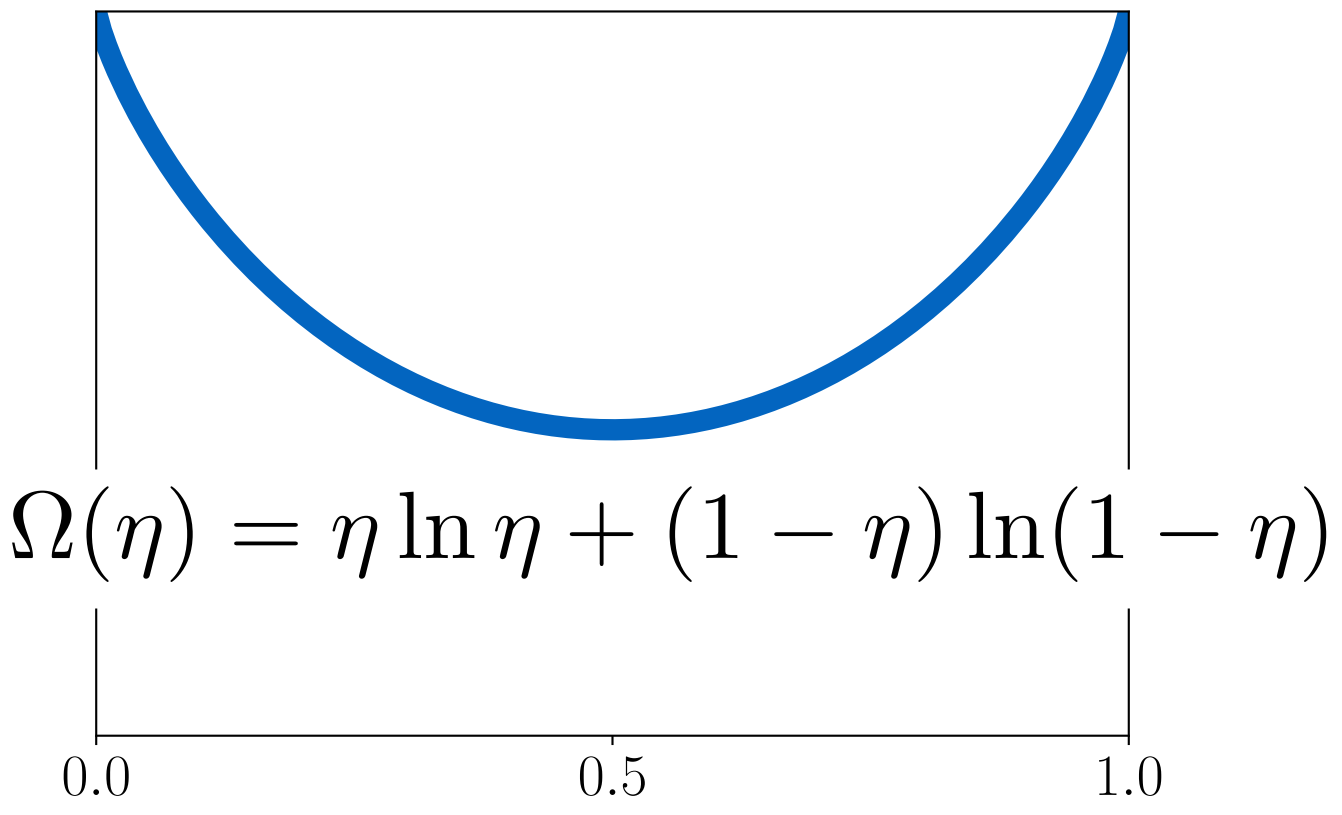


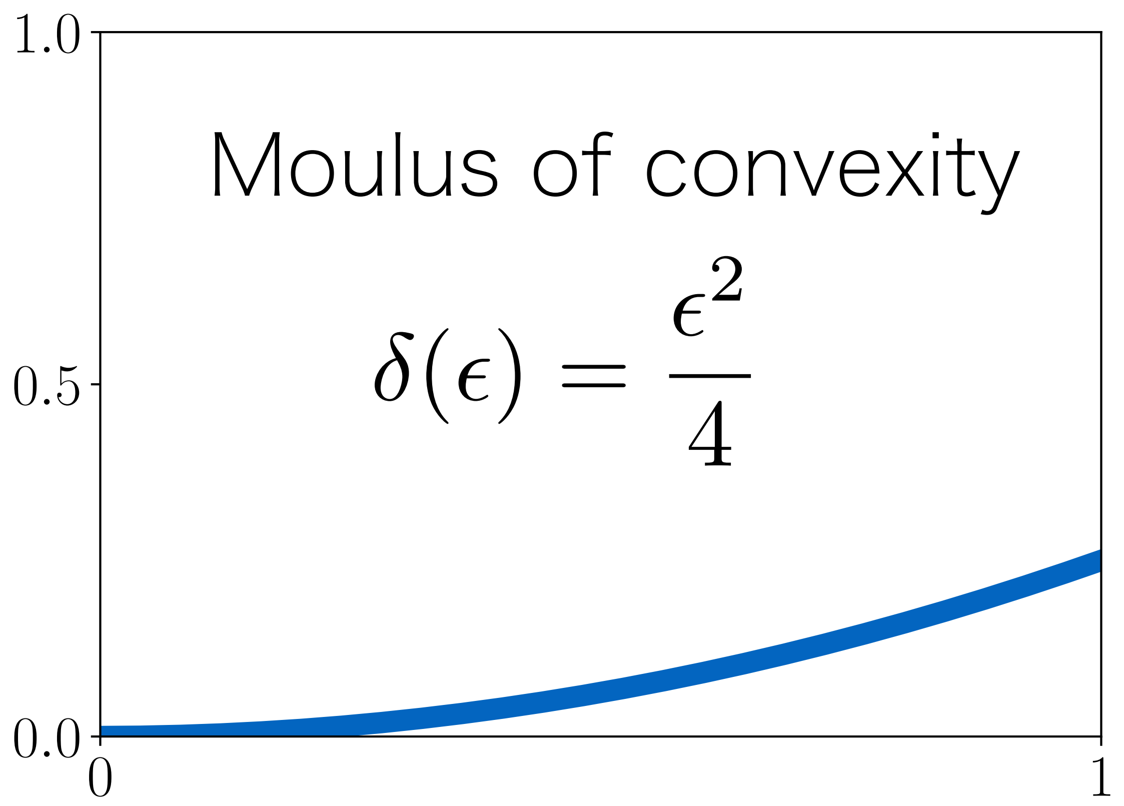
$(\Omega = -\underline{L}_\ell)$

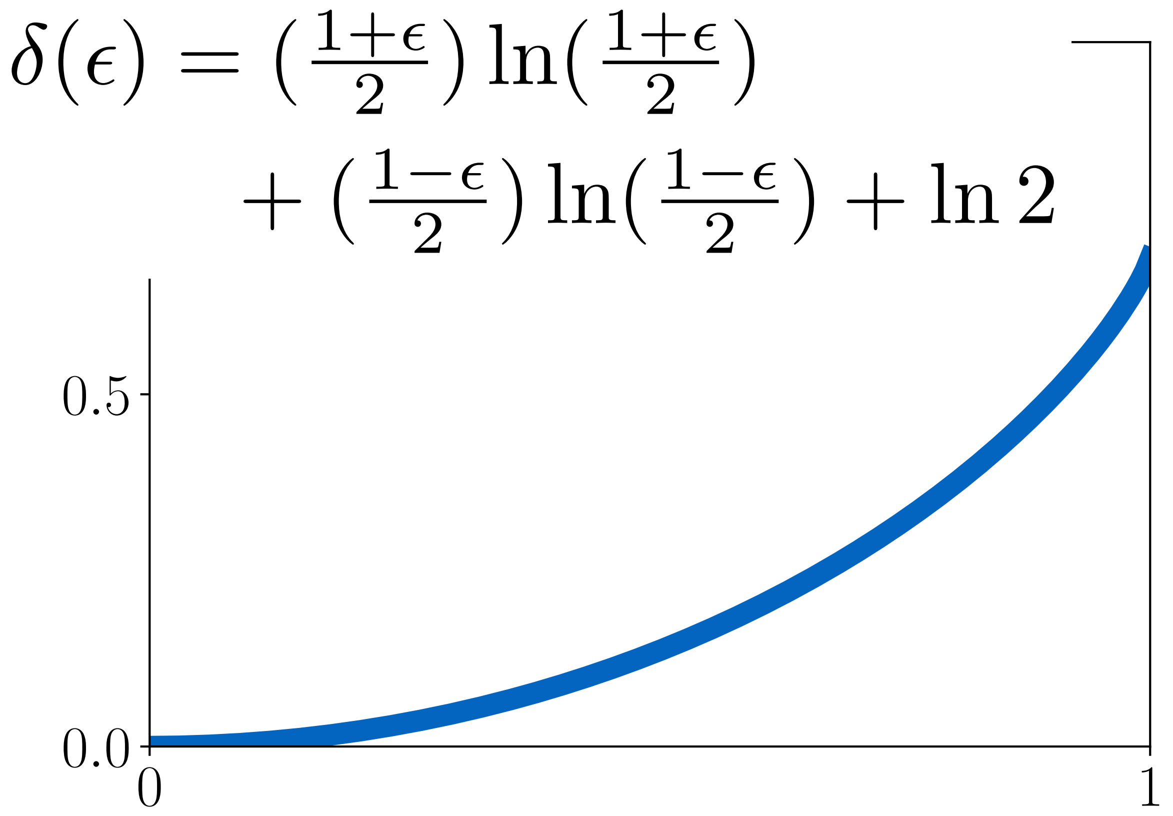
Examples

- L2 loss
- Log loss





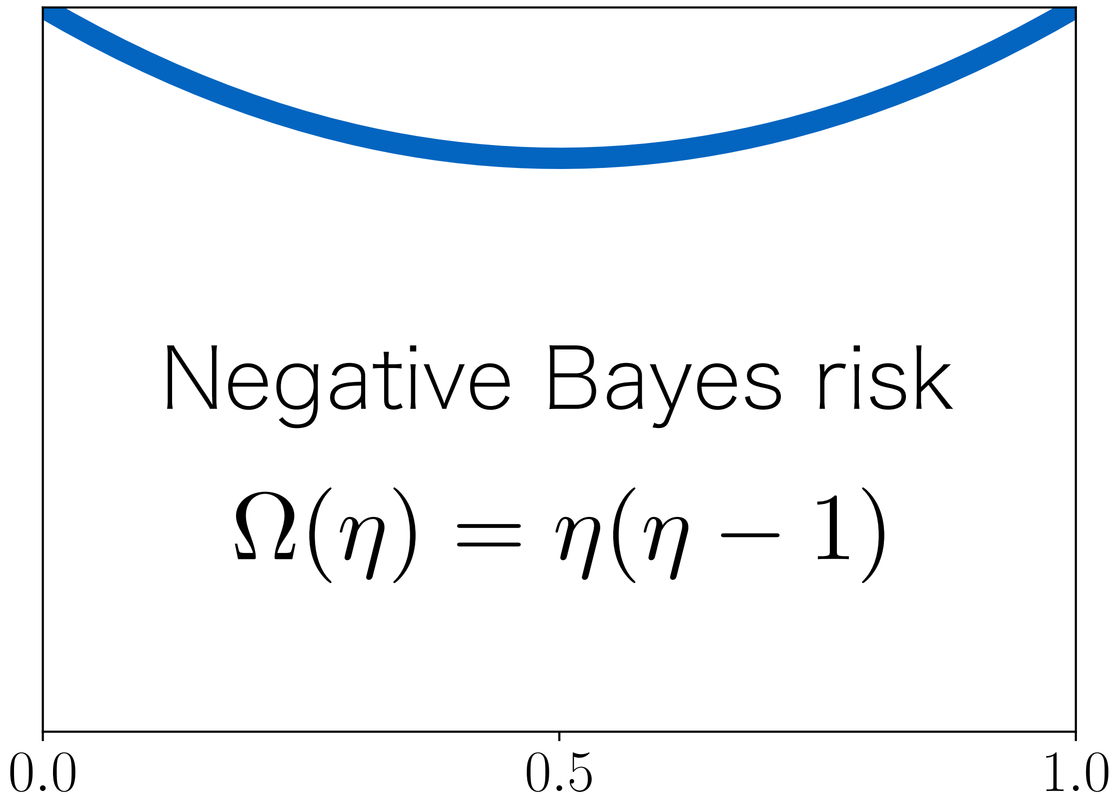


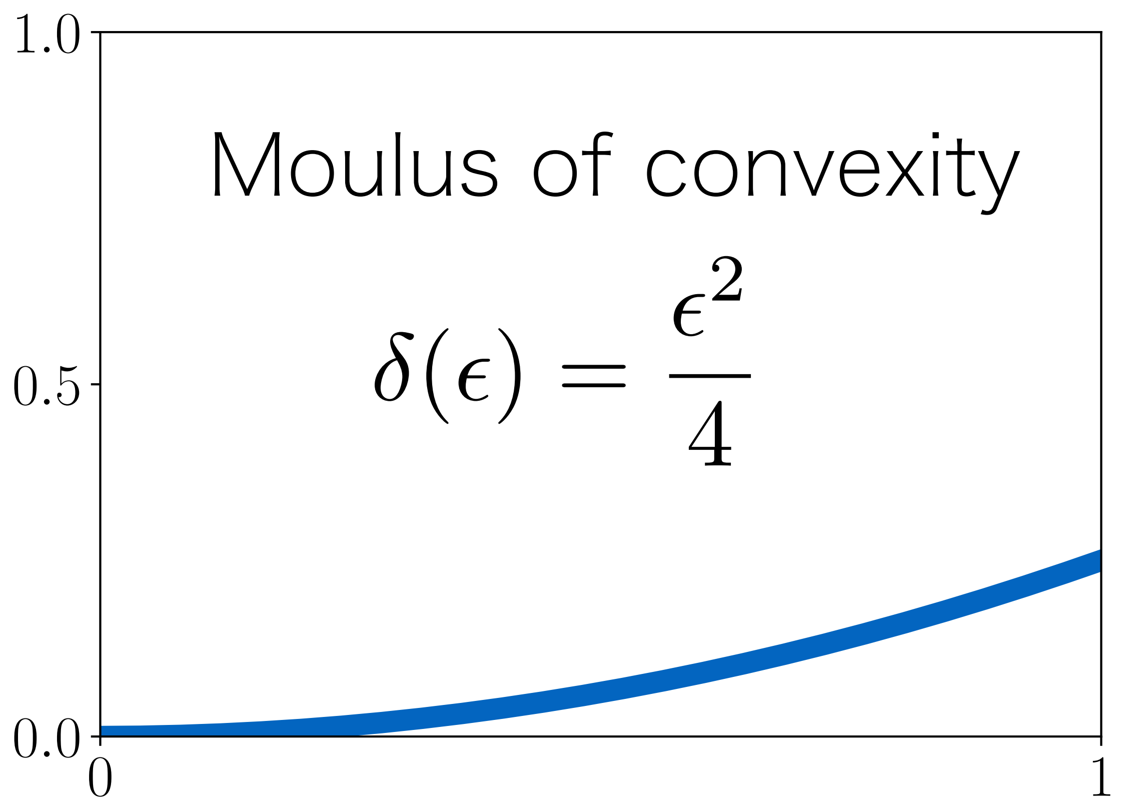


$(\Omega = -\underline{L}_\ell)$

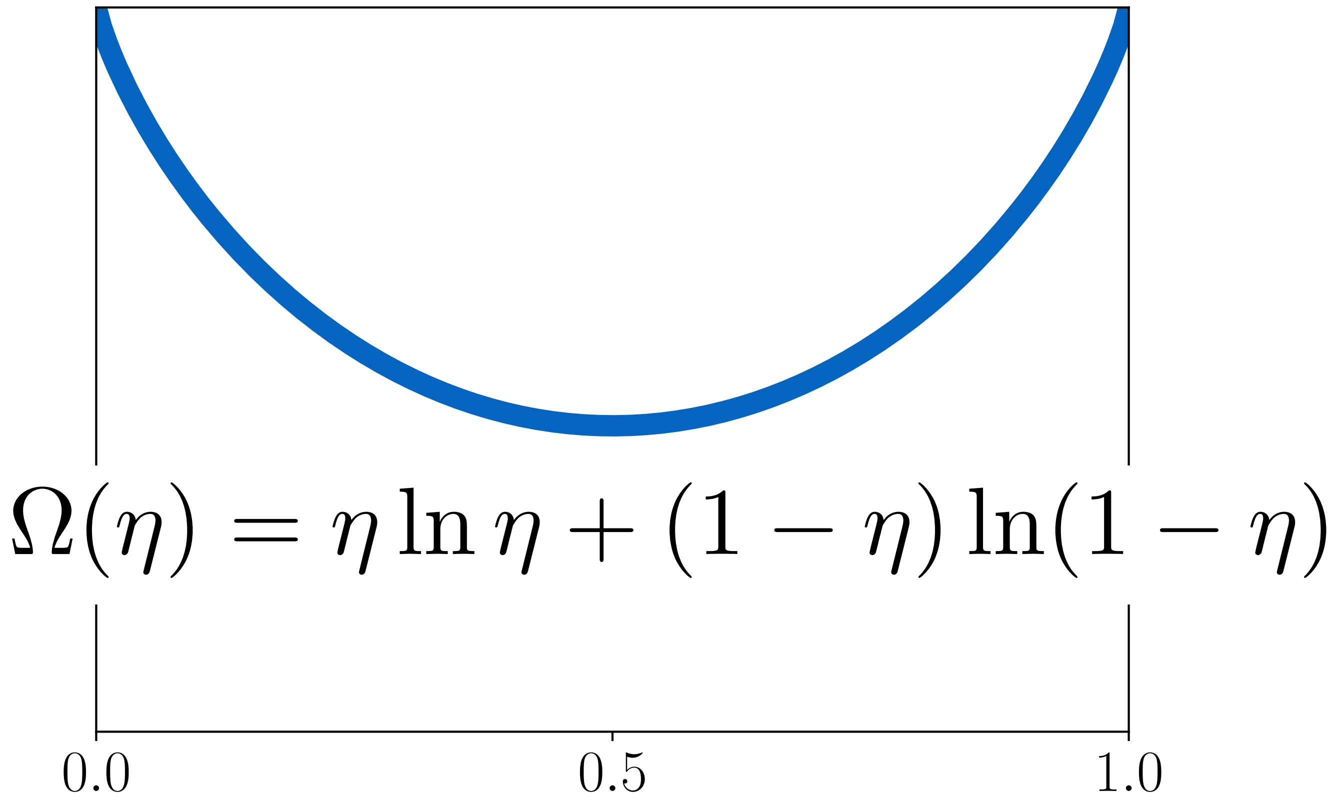
Examples

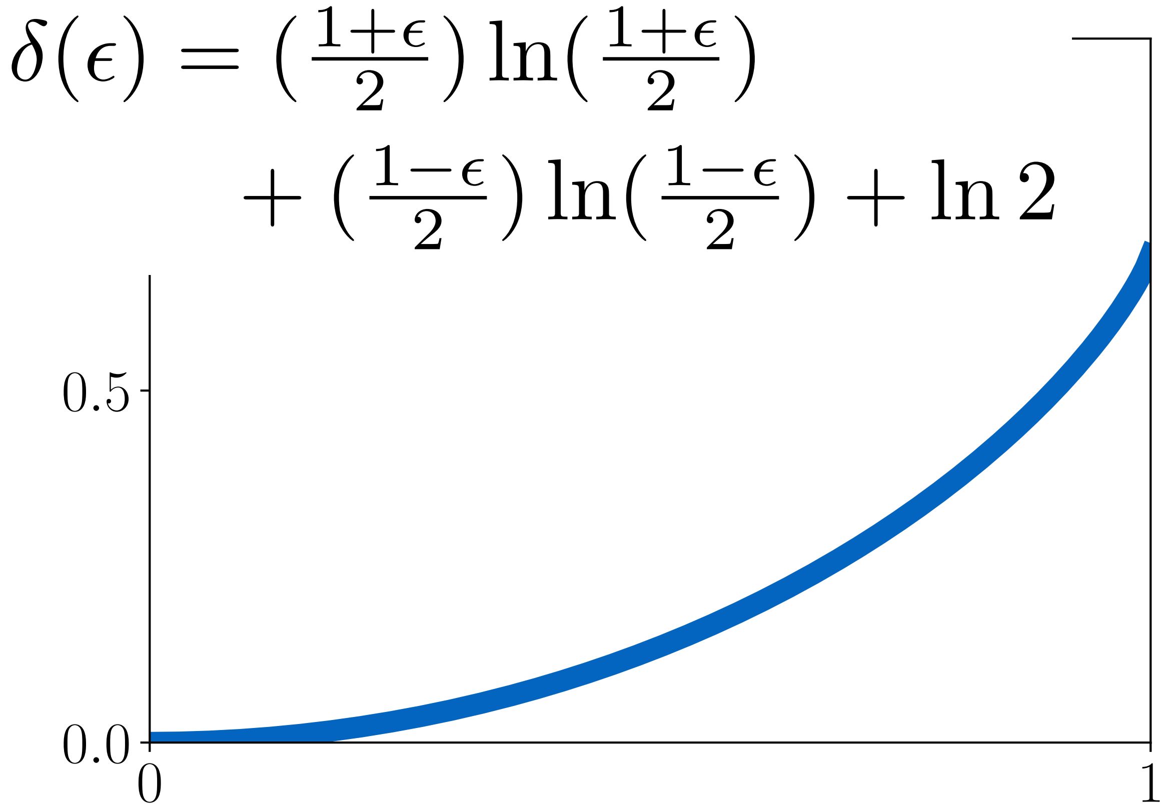
- L2 loss



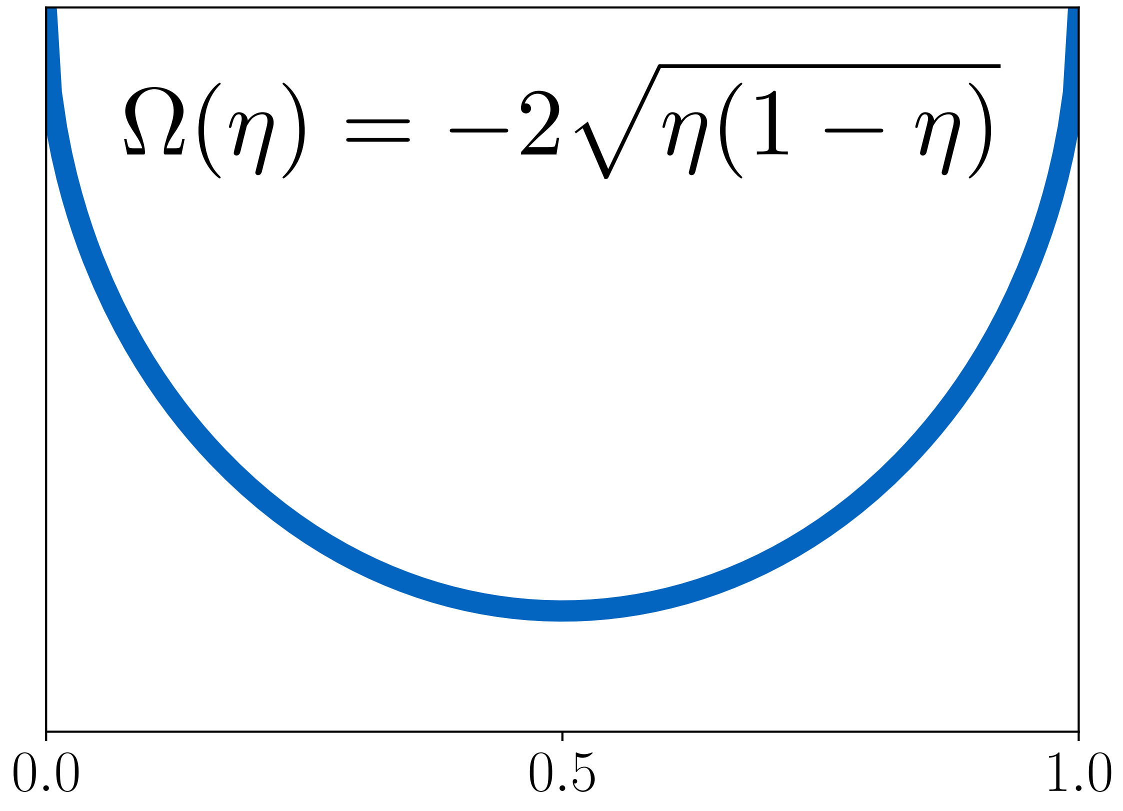


- Log loss





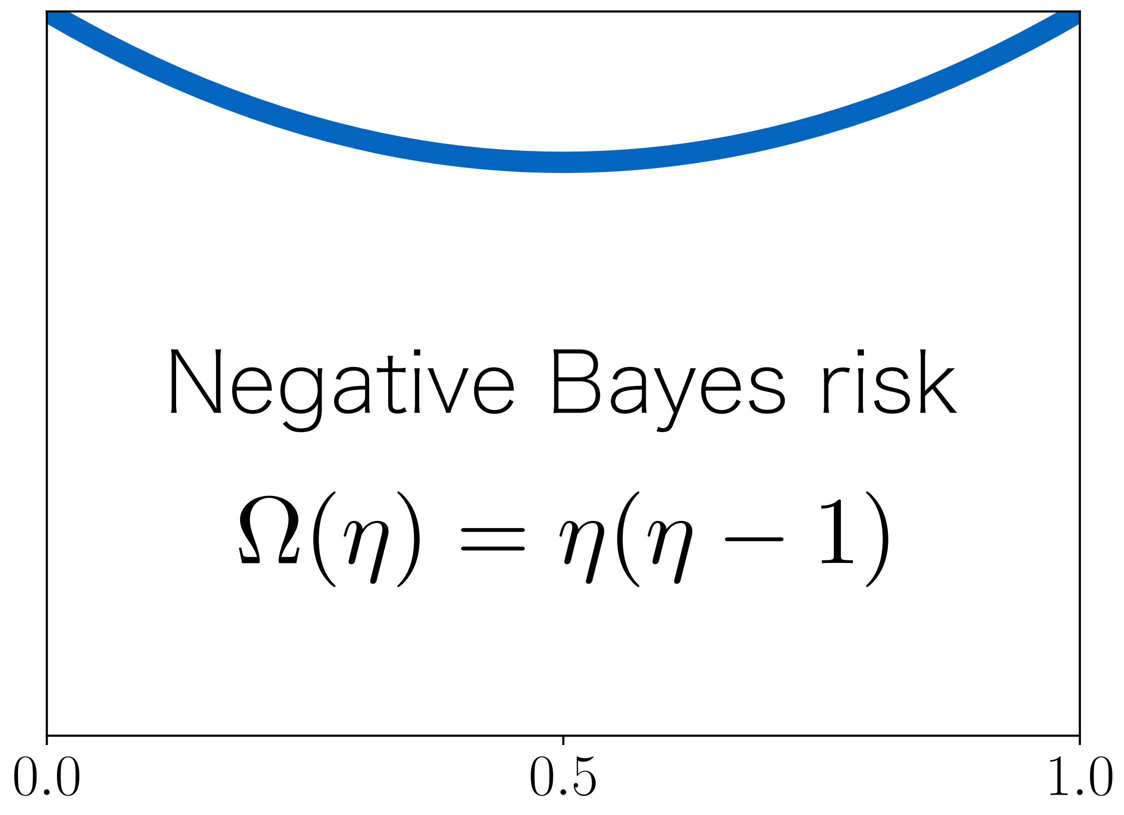
- Exponential (Boosting) loss

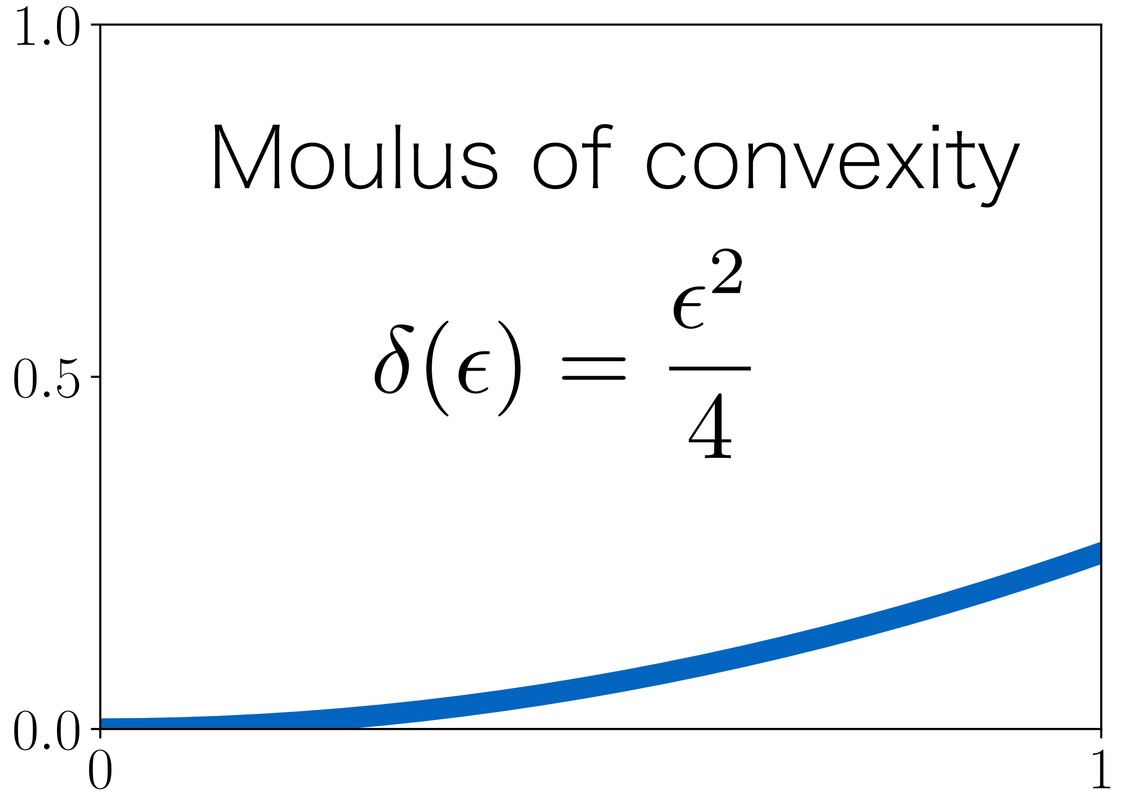


$(\Omega = -\underline{L}_\ell)$

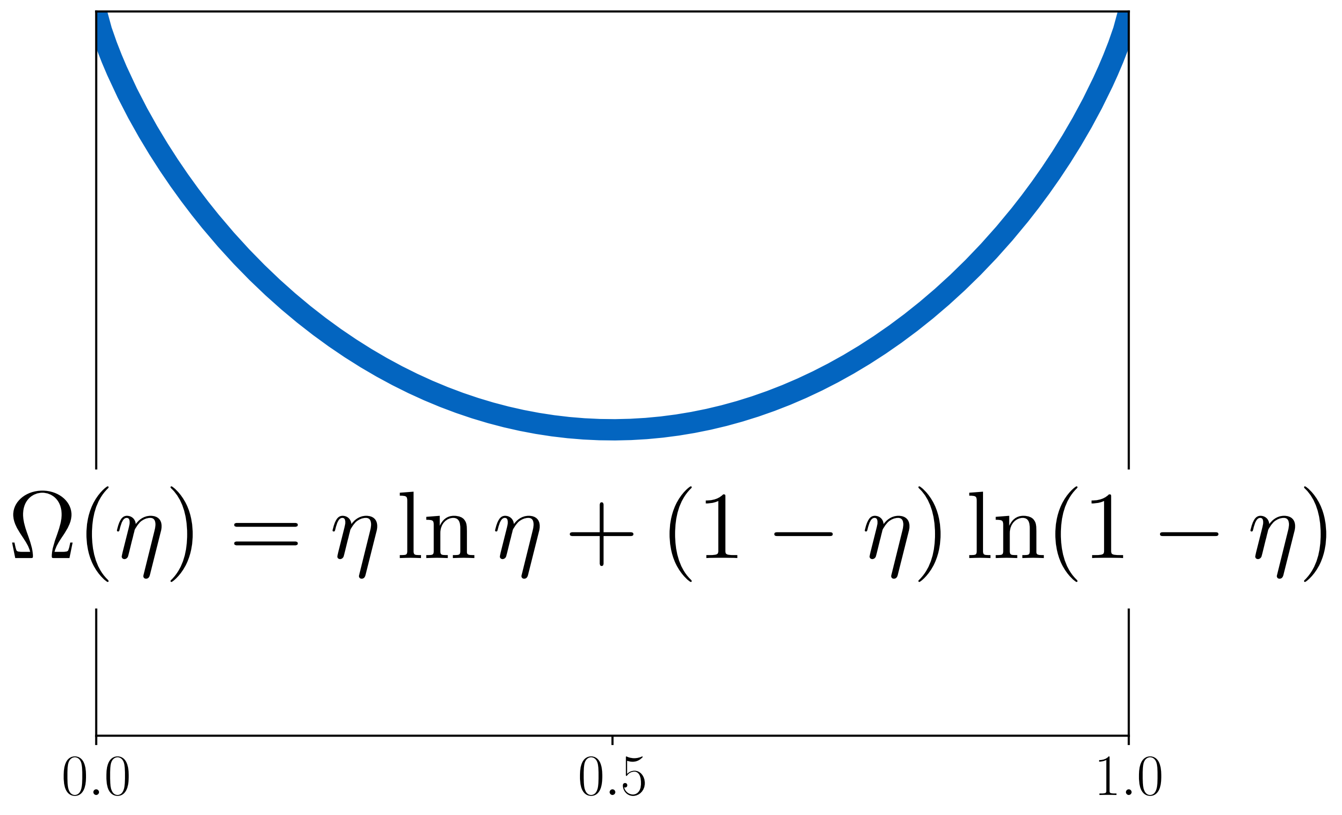
Examples

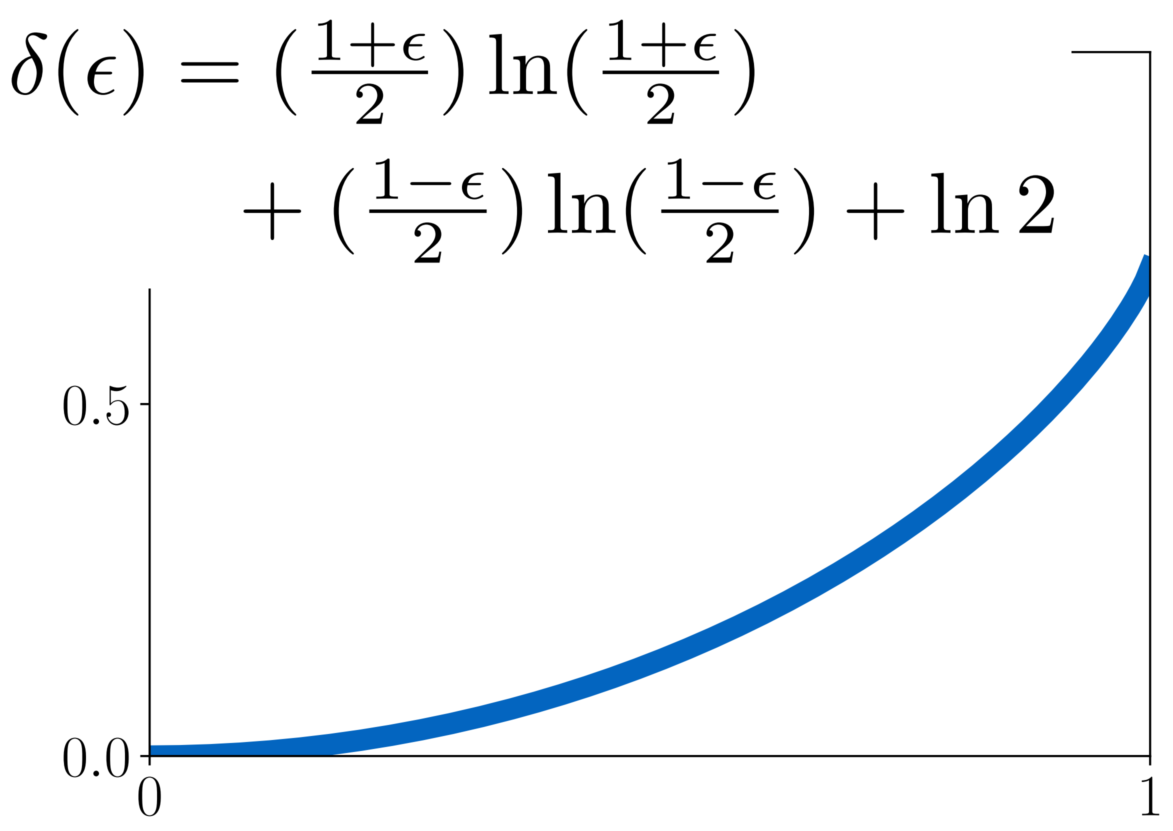
- L2 loss



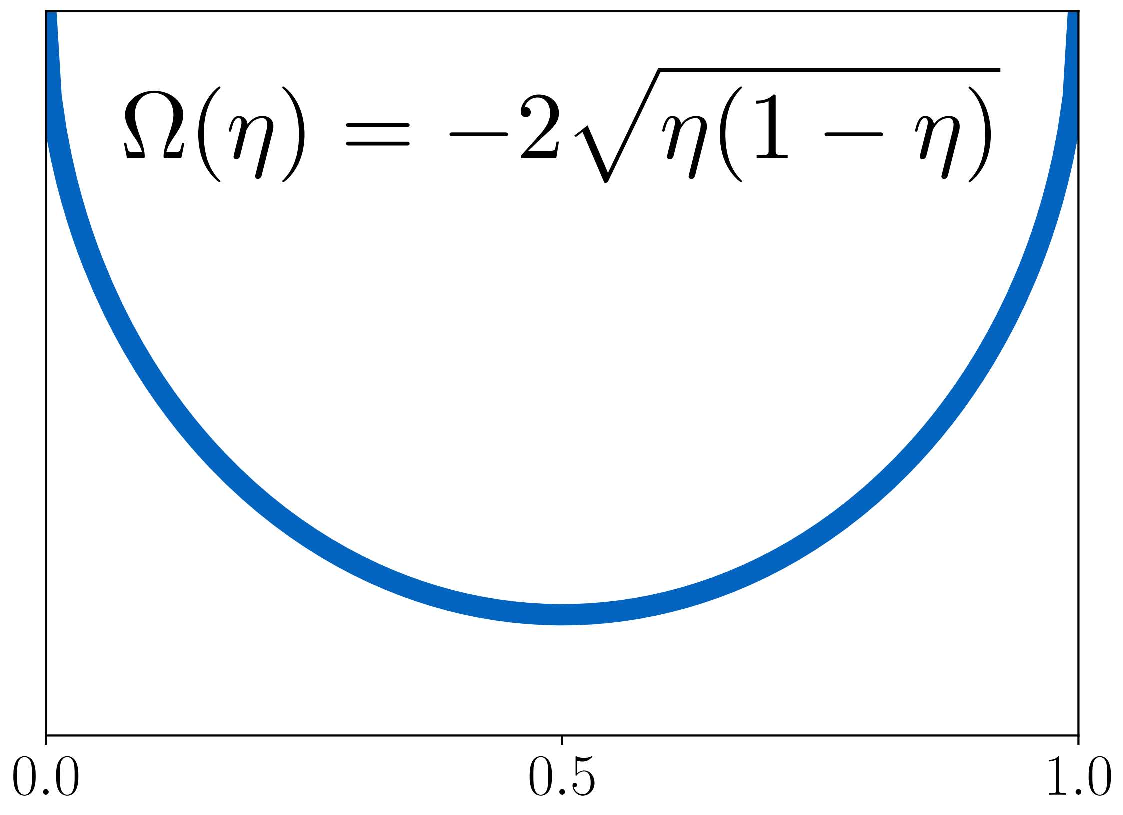


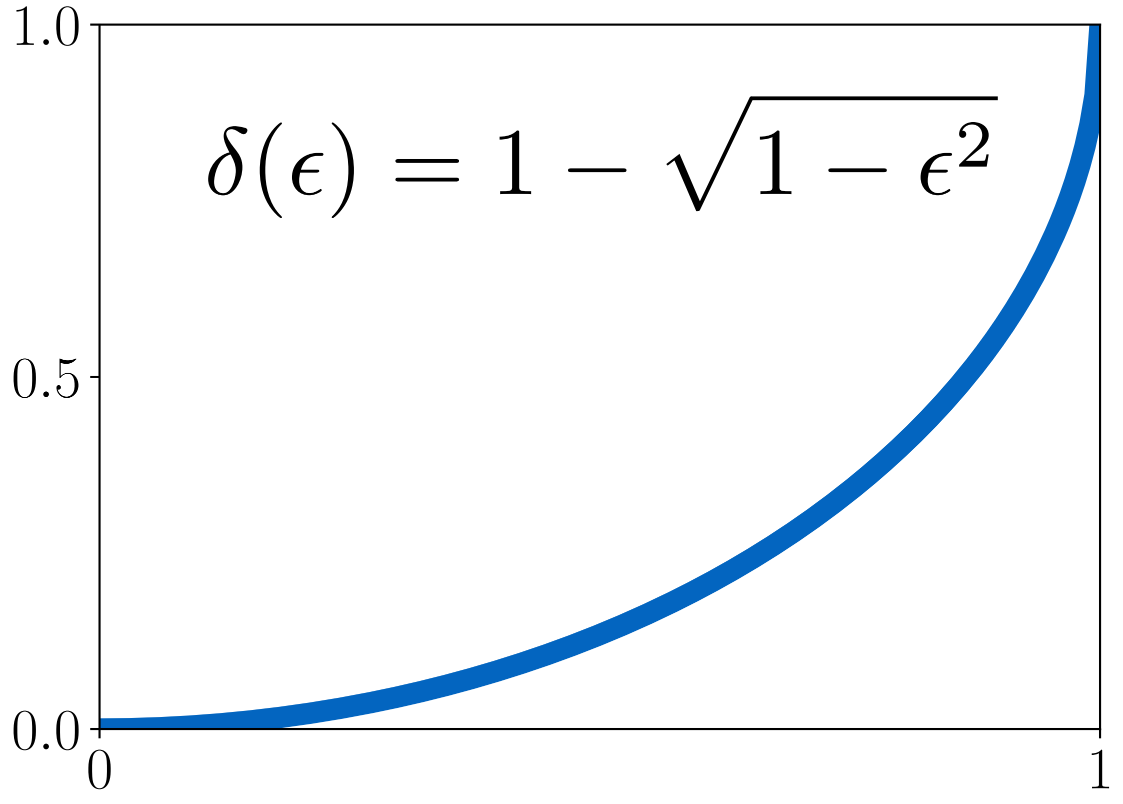
- Log loss





- Exponential (Boosting) loss

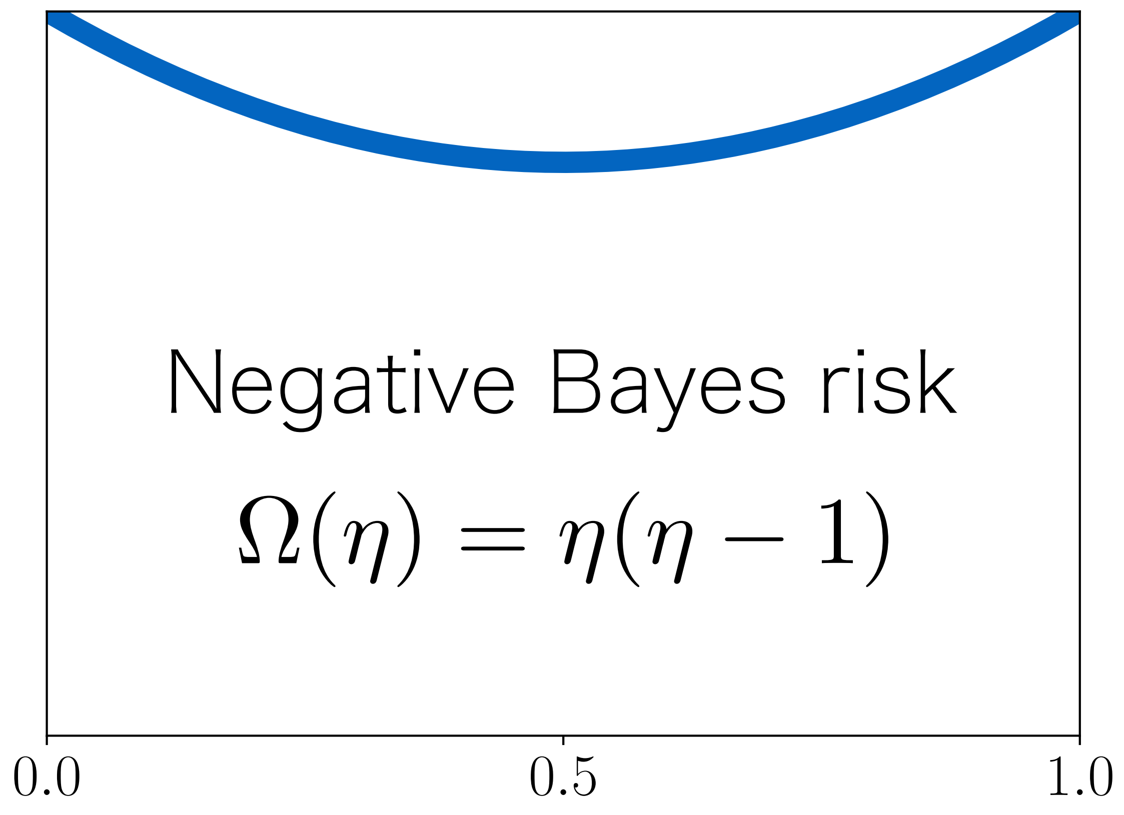


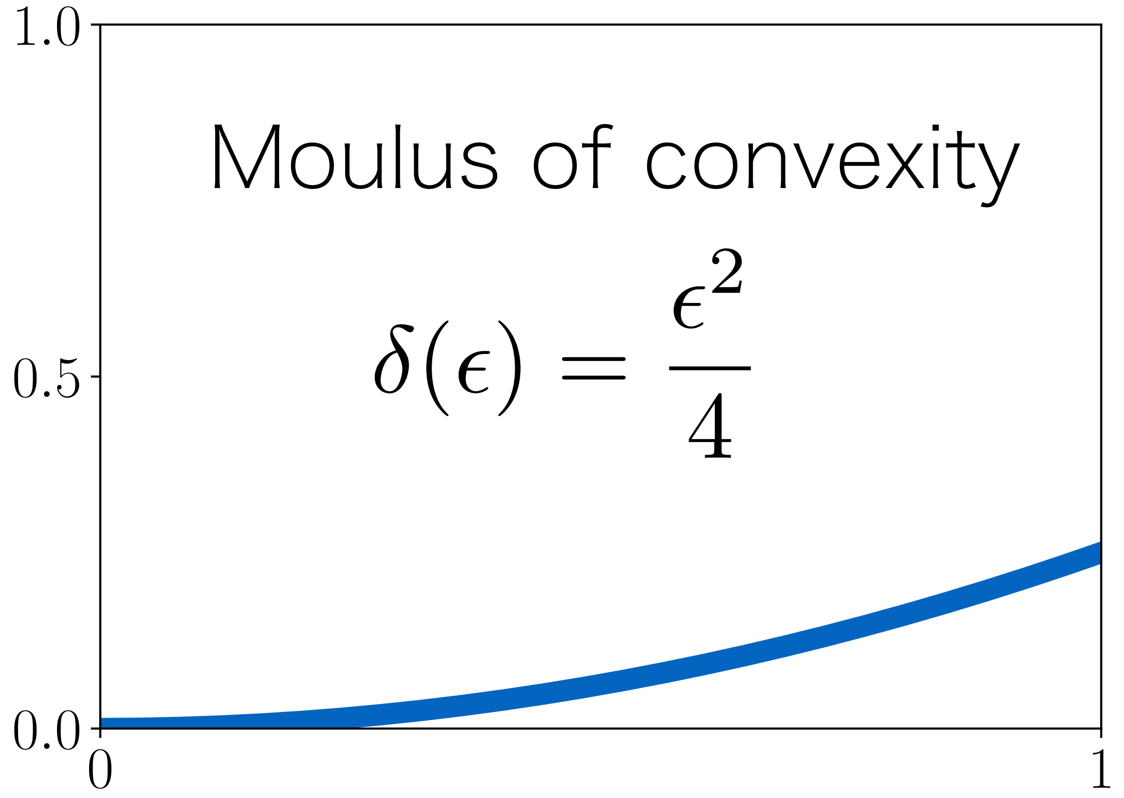


$(\Omega = -\underline{L}_\ell)$

Examples

- L2 loss





- Log loss
- Exponential (Boosting) loss

Theorem. For a proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, for all $\eta, \hat{\eta} \in [0,1]$,

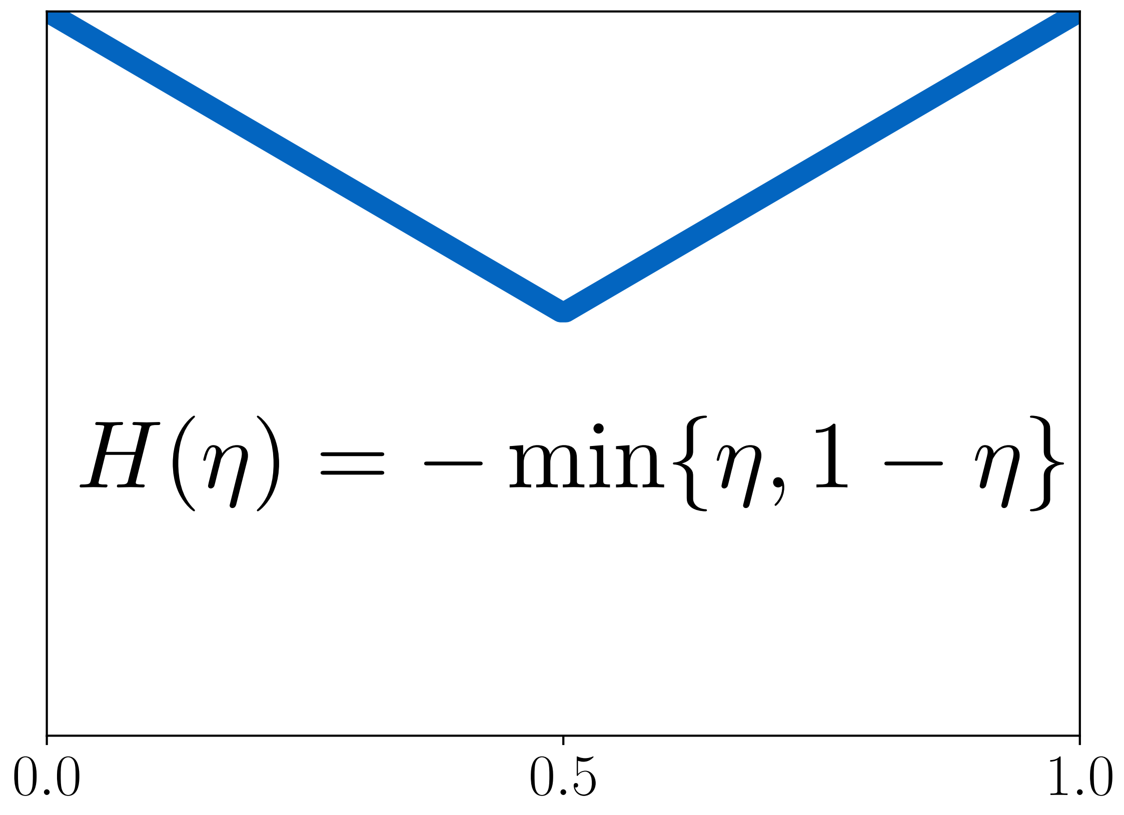
$$\delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|) \leq R_\ell(\eta, \hat{\eta}).$$

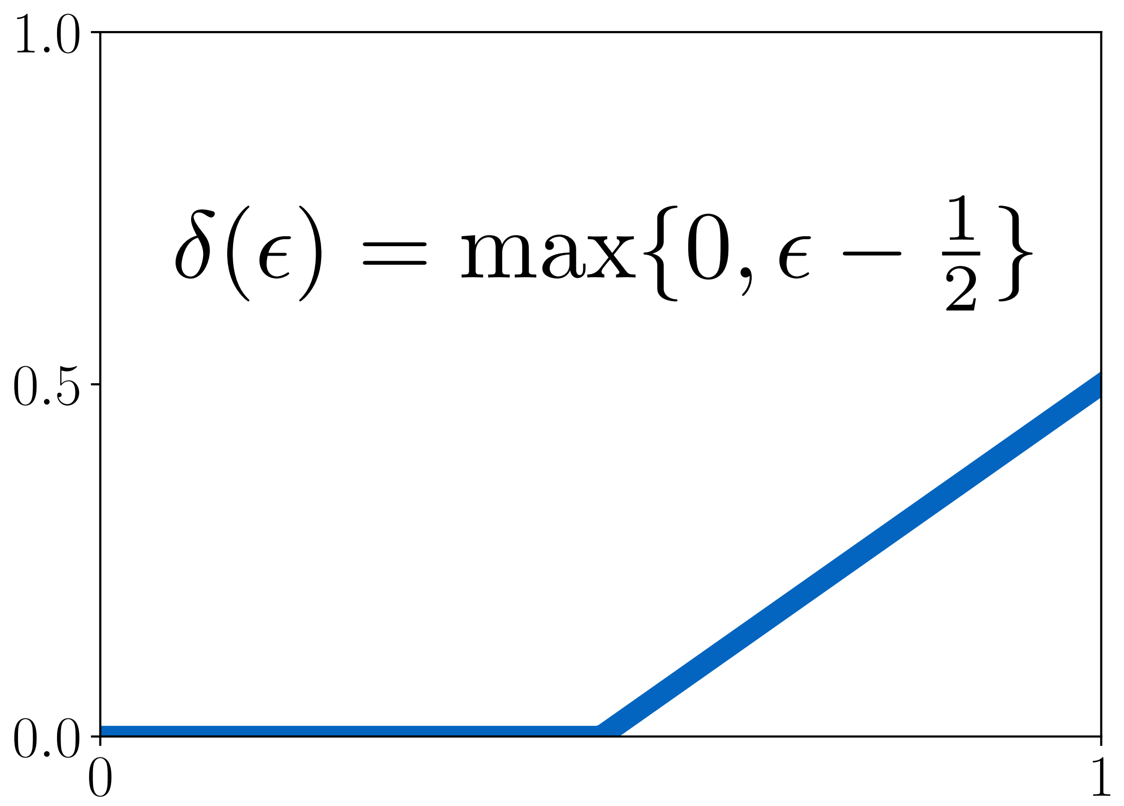
$$\frac{1}{4}|\eta - \hat{\eta}|^2 \leq R_\ell(\eta, \hat{\eta}) \implies |\eta - \hat{\eta}| \leq \sqrt{4R_\ell(\eta, \hat{\eta})}$$

$(\Omega = -\underline{L}_\ell)$

Examples

- L1 loss





Theorem. For a proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, for all $\eta, \hat{\eta} \in [0,1]$,

$$\delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|) \leq R_\ell(\eta, \hat{\eta}).$$

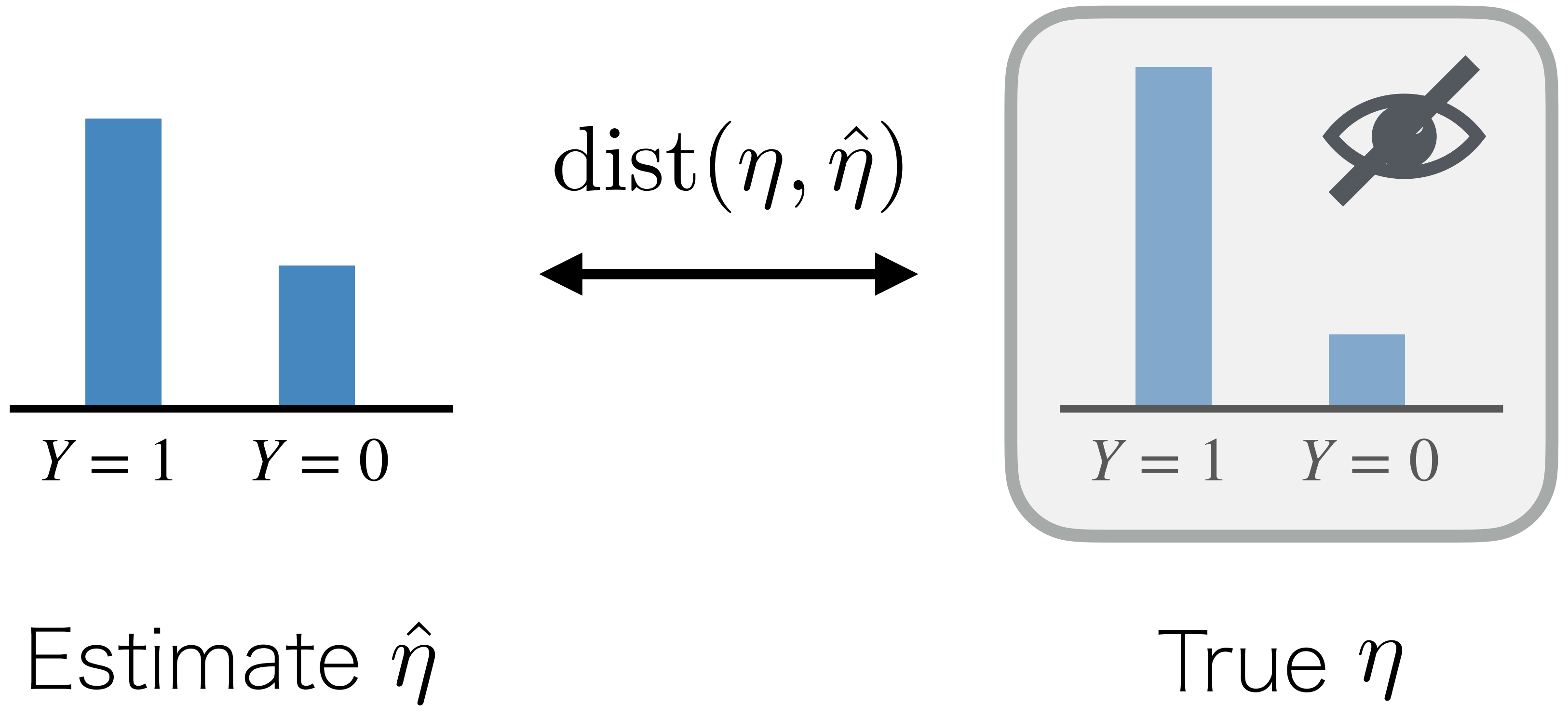
$$\max\left\{0, |\eta - \hat{\eta}| - \frac{1}{2}\right\} \leq R_\ell(\eta, \hat{\eta})$$

$$\implies \quad |\eta - \hat{\eta}| \leq R_\ell(\eta, \hat{\eta}) + \frac{1}{2}$$ (vacuous)

$(\ \Omega = -\underline{L}_\ell\ )$

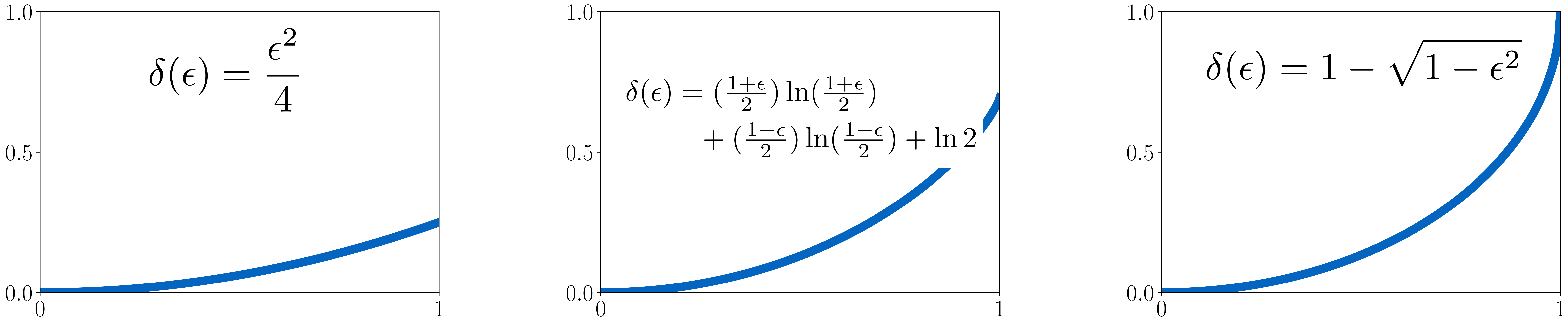
Outline

- **Q.** How should we assess probability estimates?
  - Proper losses
- **Q.** How can estimated probabilities be used for other tasks?
  - Regret bounds
- **Q.** How to compare different loss functions?
  - Order function of moduli



Can we obtain more meaningful bounds?

- L1 regret bounds are characterized by moduli



  - ❖ Moduli are not friendly for us

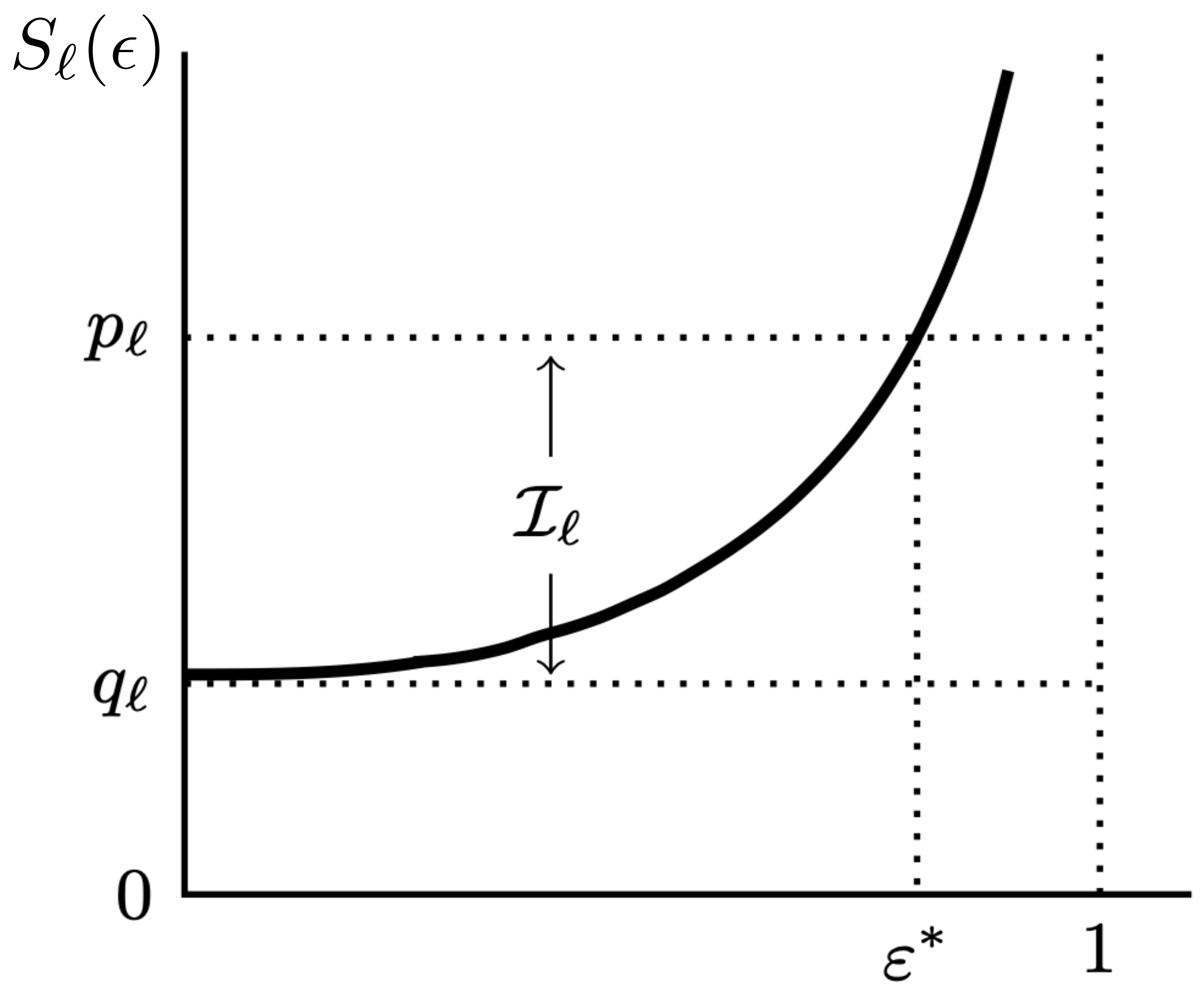
- **Q.** Can we evaluate moduli by polynomials?
  - ❖ For some r, R, $\epsilon^r \leq \delta(\epsilon) \leq \epsilon^R$

Polynomial bounds of moduli

Definition [Simonenko 1964]**.** For a proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, order function $S_\ell : (0,1] \to \overline{\mathbb{R}}$ is

$$S_\ell(t) := \frac{t(\delta^{\star\star}_{-\underline{L}_\ell})'(t)}{\delta^{\star\star}_{-\underline{L}_\ell}(t)}.$$

- Order function tells how many orders polynomial approximation of $\delta^{\star\star}_{-\underline{L}_\ell}$ is at a given point t
  - Take a point ϵ^*
  - Take sup and inf in $[0, \epsilon^*]$ to define p_ℓ and q_ℓ
  - p_ℓ and q_ℓ provides us polynomial bounds and their orders (see next page)



Example for log loss

Igor Borisovich Simonenko. Interpolation and extrapolation of linear operators in Orlicz spaces. *Matematicheskii Sbornik*, 105(4):536–553, 1964.

Polynomial bounds of moduli

Definition [Simonenko 1964]**.** For a proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, order function $S_\ell : (0,1] \to \overline{\mathbb{R}}$ is

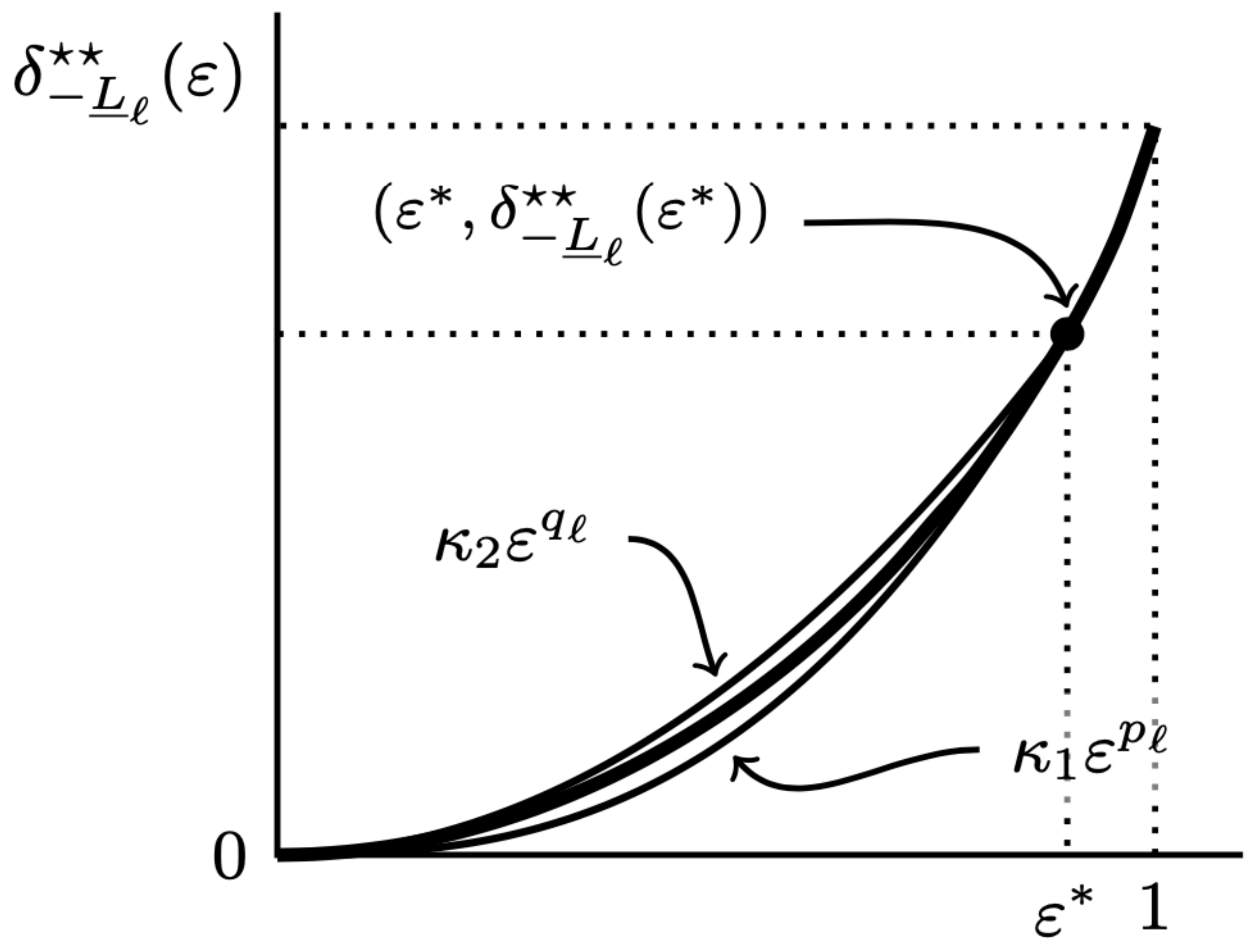
$$S_\ell(t) := \frac{t(\delta^{\star\star}_{-\underline{L}_\ell})'(t)}{\delta^{\star\star}_{-\underline{L}_\ell}(t)}.$$

Theorem. For a strictly proper loss $\ell : \{0,1\} \times [0,1] \to \mathbb{R}_{\geq 0}$, $\epsilon_* \in (0,1]$, define $p_\ell := \sup_{t \in (0,\epsilon_*]} S_\ell(t)$ and $q_\ell := \inf_{t \in (0,\epsilon_*]} S_\ell(t)$.

Then, for all $\epsilon \in [0, \epsilon_*]$, we have

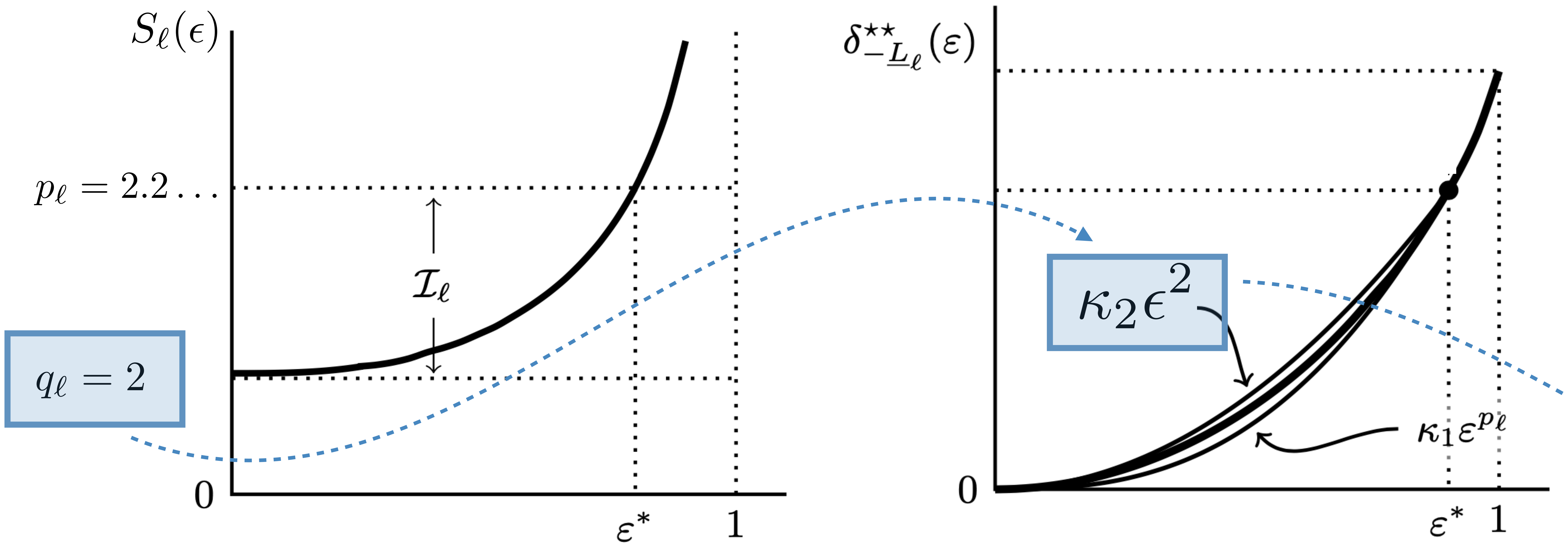
$$\kappa_1 \epsilon^{p_\ell} \leq \delta^{\star\star}_{-\underline{L}_\ell}(\epsilon) \leq \kappa_2 \epsilon^{q_\ell},$$

where $\kappa_1 := \frac{\delta^{\star\star}_{-\underline{L}_\ell}(\epsilon_*)}{\epsilon_*^{p_\ell}}$ and $\kappa_2 := \frac{\delta^{\star\star}_{-\underline{L}_\ell}(\epsilon_*)}{\epsilon_*^{q_\ell}}$.



Examples

- Log loss $\delta(\epsilon) = (\frac{1+\epsilon}{2})\ln(\frac{1+\epsilon}{2}) + (\frac{1-\epsilon}{2})\ln(\frac{1-\epsilon}{2}) + \ln 2$



$$|\eta - \hat{\eta}| \leq R_\ell(\eta, \hat{\eta})^{\frac{1}{2.2\ldots}}$$

Upper bound cannot be better than $R_\ell(\eta, \hat{\eta})^2$

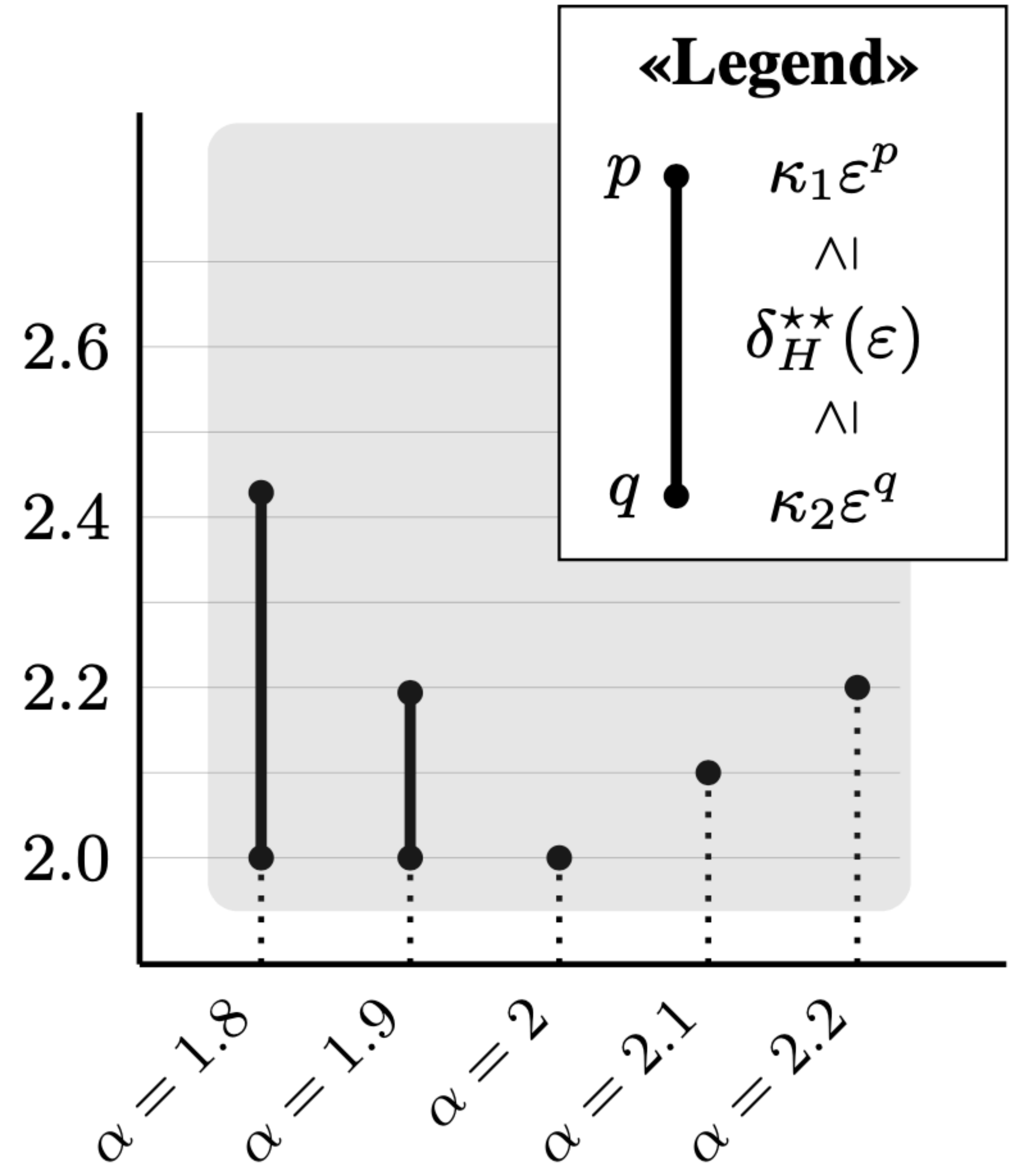
Order function

Modulus $\delta(\epsilon) = O(\epsilon^2)$

$\delta(\epsilon) = \Omega(\epsilon^{2.2\ldots})$

Examples

- Polynomial entropies $\Omega(\eta) = |\eta - \frac{1}{2}|^\alpha - \frac{1}{2^\alpha}$
- It reduces to (the associated Bayes risk of) L2 loss when $\alpha = 2$



- Implications
  - No matter how we modulate α, the smallest q_ℓ is 2
  - The upper order p_ℓ is tight when $\alpha = 2$ and we obtain $|\eta - \hat{\eta}| \leq R_\ell(\eta, \hat{\eta})^2$

Reminder

- [⇐] For a concave $H : [0,1] \to \mathbb{R}$, loss $\ell(y, \hat{\eta}) = H(\hat{\eta}) + (y - \hat{\eta})H'(\hat{\eta})$ is proper
  - Remark: one-to-one correspondence between proper loss and concave function

Summary

- **Proper loss**: a reasonable loss for probabilistic estimation

Definition. $\ell(y, \hat{\eta})$ is strictly proper iff $L_\ell(\eta, \hat{\eta}) = \underline{L}_\ell(\eta) \iff \hat{\eta} = \eta$ for all $\eta \in [0, 1]$.

$$L_\ell(\eta, \hat{\eta}) = \eta\ell(1, \hat{\eta}) + (1 - \eta)\ell(0, \hat{\eta}) \qquad \underline{L}_\ell(\eta) = \inf_{\hat{\eta}\in[0,1]} L_\ell(\eta, \hat{\eta})$$

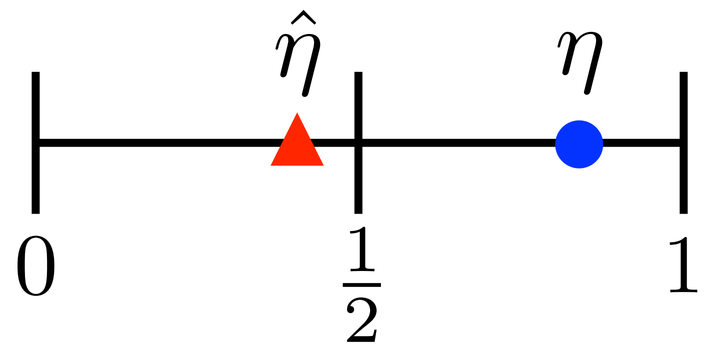
- L1 **regret bound** is characterized by **modulus of convexity**

Theorem. For a proper loss $\ell : \{0, 1\} \times [0, 1] \to \mathbb{R}_{\geq 0}$, for all $\eta, \hat{\eta} \in [0, 1]$,

$$\delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|) \leq R_\ell(\eta, \hat{\eta}).$$

❖ Useful for unifying regret bounds of many downstream tasks

Classification



Ranking