Following the Flattened Leader

Wojciech Kotłowski[1] Peter Grünwald[1] Steven de Rooij[2]

[1]National Research Institute for Mathematics and Computer Science (CWI)
The Netherlands

[2]University of Cambridge

COLT 2010

Outline

1. Sequential prediction with log-loss.
   - Set of experts = exponential family.

Outline

1. Sequential prediction with log-loss.
   - Set of experts = exponential family.
2. Prediction strategies:

Bayes strategy:
- ☺ achieves optimal regret
- ☹ usually hard to calculate

"Follow the leader" strategy:
- ☺ simple to compute/update
- ☹ suboptimal

Outline

1. Sequential prediction with log-loss.
   - Set of experts = exponential family.
2. Prediction strategies:

Bayes strategy:

☺ achieves optimal regret

☹ usually hard to calculate

"Follow the leader" strategy:

☺ simple to compute/update

☹ suboptimal

3. Our contribution

"Follow the flattened leader" strategy:

A slight modification of "follow the leader".

☺ achieves performance of Bayes

☺ retains simplicity of ML

Outline

1. Sequential prediction with log-loss.
   - Set of experts = exponential family.
2. Prediction strategies:

Bayes strategy:
☺ achieves optimal regret
☹ usually hard to calculate

"Follow the leader" strategy:
☺ simple to compute/update
☹ suboptimal

3. Our contribution

"Follow the flattened leader" strategy:
A slight modification of "follow the leader".
☺ achieves performance of Bayes
☺ retains simplicity of ML

4. Applications: prediction, coding, model selection.

Sequential Prediction

- Family of distributions (model) $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$.

Sequential Prediction

- Family of distributions (model) $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$.
- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}^\infty$, revealed one by one.

Sequential Prediction

- Family of distributions (model) $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$.
- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}^\infty$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, predict x_{n+1} by assigning a distribution $P(\cdot | x^n)$.

Sequential Prediction

- Family of distributions (model) $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$.
- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}^\infty$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, predict x_{n+1} by assigning a distribution $P(\cdot | x^n)$.
- After x_{n+1} is revealed, incur log-loss $-\log P(x_{n+1} | x^n)$.

Sequential Prediction

- Family of distributions (model) $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$.
- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}^\infty$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, predict x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, incur log-loss $-\log P(x_{n+1}|x^n)$.
- Regret w.r.t. the best "expert" from $\mathcal{M}$:

$$\mathcal{R}(P, x^n) = \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\mu \in \Theta} \sum_{i=1}^{n} -\log P_\mu(x_i|x^{i-1}).$$

Sequential Prediction

- Family of distributions (model) $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$.
- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}^\infty$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, predict x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, incur log-loss $-\log P(x_{n+1}|x^n)$.
- Regret w.r.t. the best "expert" from $\mathcal{M}$:

$$\mathcal{R}(P, x^n) = \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\mu\in\Theta} \sum_{i=1}^{n} -\log P_\mu(x_i|x^{i-1}).$$

- Process generating the outcomes:
  - adversarial: only boundedness assumptions on x^n,
  - stochastic: $X_1, X_2, \ldots$ i.i.d. $\sim P^*$, possibly $P^* \notin \mathcal{M}$, $\mathcal{R}(P, X^n)$ is a random variable.

Sequential Prediction: Example

- $\mathcal{M} = \{P_\mu | \mu \in [0,1]\}$, P_μ Bernoulli.
- $x^n = 1010110110.$

Sequential Prediction: Example

- $\mathcal{M} = \{P_\mu | \mu \in [0,1]\}$, P_μ Bernoulli.
- $x^n = 1010110110$.
- Best expert in $\mathcal{M}$: $P_{\hat{\mu}_n}$, $\hat{\mu}_n = \frac{\#1}{n}$ $(= \frac{3}{5})$.

Sequential Prediction: Example

- $\mathcal{M} = \{P_\mu | \mu \in [0,1]\}$, P_μ Bernoulli.
- $x^n = 1010110110$.
- Best expert in $\mathcal{M}$: $P_{\hat{\mu}_n}$, $\hat{\mu}_n = \frac{\#1}{n}$ $(= \frac{3}{5})$.
- "Follow the leader" prediction strategy:
  - $P(\cdot|x^i) = P_{\hat{\mu}_i}(\cdot)$

Sequential Prediction: Example

- $\mathcal{M} = \{P_\mu | \mu \in [0,1]\}$, P_μ Bernoulli.
- $x^n = 1010110110$.
- Best expert in $\mathcal{M}$: $P_{\hat{\mu}_n}$, $\hat{\mu}_n = \frac{\#1}{n}$ $(= \frac{3}{5})$.
- "Follow the leader" prediction strategy:
  - $P(\cdot|x^i) = P_{\hat{\mu}_i}(\cdot) \Longleftarrow \hat{\mu}_0$ undefined, $P(x_2|x_1) = 0\ldots$

Sequential Prediction: Example

- $\mathcal{M} = \{P_\mu | \mu \in [0,1]\}$, P_μ Bernoulli.
- $x^n = 1010110110.$
- Best expert in $\mathcal{M}$: $P_{\hat{\mu}_n}$, $\hat{\mu}_n = \frac{\#1}{n}$ $(= \frac{3}{5})$.
- "Follow the leader" prediction strategy:
  - $P(\cdot|x^i) = P_{\hat{\mu}_i}(\cdot) \Longleftarrow \hat{\mu}_0$ undefined, $P(x_2|x_1) = 0\ldots$
  - $P(\cdot|x^i) = P_{\hat{\mu}_i^\circ}(\cdot)$, $\hat{\mu}_i^\circ = \frac{\#1+1}{n+2}$ (Laplace's rule of succesion).

    $\hat{\mu}_i^\circ$: $\frac{1}{2}, \frac{2}{3}, \frac{1}{2}, \frac{3}{5}, \frac{1}{2}, \frac{4}{7}, \frac{5}{8}, \frac{5}{9}, \frac{3}{5}, \frac{7}{11}, \frac{7}{12}.$

Sequential Prediction: Example

- $\mathcal{M} = \{P_\mu | \mu \in [0,1]\}$, P_μ Bernoulli.
- $x^n = 1010110110$.
- Best expert in $\mathcal{M}$: $P_{\hat{\mu}_n}$, $\hat{\mu}_n = \frac{\#1}{n}$ $(= \frac{3}{5})$.
- "Follow the leader" prediction strategy:
  - $P(\cdot|x^i) = P_{\hat{\mu}_i}(\cdot) \Longleftarrow \hat{\mu}_0$ undefined, $P(x_2|x_1) = 0 \ldots$
  - $P(\cdot|x^i) = P_{\hat{\mu}_i^\circ}(\cdot)$, $\hat{\mu}_i^\circ = \frac{\#1+1}{n+2}$ (Laplace's rule of succesion).
    $\hat{\mu}_i^\circ$: $\frac{1}{2}$, $\frac{2}{3}$, $\frac{1}{2}$, $\frac{3}{5}$, $\frac{1}{2}$, $\frac{4}{7}$, $\frac{5}{8}$, $\frac{5}{9}$, $\frac{3}{5}$, $\frac{7}{11}$, $\frac{7}{12}$.
- If x^∞ such that for large n, $\hat{\mu}_n$ bounded away from $\{0,1\}$:

$$\mathcal{R}(P, x^n) = \frac{1}{2}\log n + O(1).$$

Problem Statement

Problem Statement

- $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$ is k-parameter exponential family
  - Bernoulli, Gaussian, Poisson, gamma, beta, geometric, χ^2, ...
  - Mean-value parametrization, $\mu = E[X]$.

Problem Statement

- $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$ is k-parameter exponential family
  - Bernoulli, Gaussian, Poisson, gamma, beta, geometric, χ^2, ...
  - Mean-value parametrization, $\mu = E[X]$.
- Bayes strategy:

$$P_{\text{BAYES}}(x_{n+1}|x^n) = \int_\Theta P_\mu(x_{n+1})\,\mathrm{d}\pi(\mu|x^n)$$

  - $P_{\text{BAYES}}(x_{n+1}|x^n) \notin \mathcal{M}$ (strategy outside model).

Problem Statement

- $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$ is k-parameter exponential family
  - Bernoulli, Gaussian, Poisson, gamma, beta, geometric, χ^2, ...
  - Mean-value parametrization, $\mu = E[X]$.
- Bayes strategy:

$$P_{\text{BAYES}}(x_{n+1}|x^n) = \int_\Theta P_\mu(x_{n+1})\,\mathrm{d}\pi(\mu|x^n)$$

  - $P_{\text{BAYES}}(x_{n+1}|x^n) \notin \mathcal{M}$ (strategy outside model).
  - $\mathcal{R}(P_{\text{BAYES}}, x^n) = \frac{k}{2}\log n + O(1)$ (asympt. optimal).

Problem Statement

- $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$ is k-parameter exponential family
  - Bernoulli, Gaussian, Poisson, gamma, beta, geometric, χ^2, ...
  - Mean-value parametrization, $\mu = E[X]$.
- Bayes strategy:

$$P_{\text{BAYES}}(x_{n+1}|x^n) = \int_\Theta P_\mu(x_{n+1})\, \mathrm{d}\pi(\mu|x^n)$$

  - $P_{\text{BAYES}}(x_{n+1}|x^n) \notin \mathcal{M}$ (strategy outside model).
  - $\mathcal{R}(P_{\text{BAYES}}, x^n) = \frac{k}{2}\log n + O(1)$ (asympt. optimal).
- Plug-in strategy:

$$P_{\text{PLUG-IN}}(x_{n+1} \mid x^n) = P_{\bar{\mu}(x^n)}(x_{n+1}), \quad \bar{\mu}\colon \mathcal{X}^\infty \to \Theta$$

  - $U_{\text{PLUG-IN}}(x_{n+1} \mid x^n) \in \mathcal{M}$ (in-model strategy).

Problem Statement

- $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$ is k-parameter exponential family
  - Bernoulli, Gaussian, Poisson, gamma, beta, geometric, χ^2, ...
  - Mean-value parametrization, $\mu = E[X]$.
- Bayes strategy:

$$P_{\text{BAYES}}(x_{n+1}|x^n) = \int_\Theta P_\mu(x_{n+1})\,\mathrm{d}\pi(\mu|x^n)$$

  - $P_{\text{BAYES}}(x_{n+1}|x^n) \notin \mathcal{M}$ (strategy outside model).
  - $\mathcal{R}(P_{\text{BAYES}}, x^n) = \frac{k}{2}\log n + O(1)$ (asympt. optimal).
- Plug-in strategy:

$$P_{\text{PLUG-IN}}(x_{n+1} \mid x^n) = P_{\bar{\mu}(x^n)}(x_{n+1}), \quad \bar{\mu}\colon \mathcal{X}^\infty \to \Theta$$

  - $U_{\text{PLUG-IN}}(x_{n+1} \mid x^n) \in \mathcal{M}$ (in-model strategy).
  - ML plug-in strategy ("follow the leader") if $\bar{\mu}(x^n) = \hat{\mu}_n^\circ$:

$$\hat{\mu}_n^\circ = \frac{n_0 x_0 + \sum_{i=1}^n x_i}{n_0 + n} \qquad \text{(smoothed ML estimator)}$$

Problem Statement

- $\mathcal{M} = \{P_\mu | \mu \in \Theta\}$ is k-parameter exponential family
  - Bernoulli, Gaussian, Poisson, gamma, beta, geometric, χ^2, ...
  - Mean-value parametrization, $\mu = E[X]$.
- Bayes strategy:

$$P_{\text{BAYES}}(x_{n+1}|x^n) = \int_\Theta P_\mu(x_{n+1}) \, \mathrm{d}\pi(\mu|x^n)$$

  - $P_{\text{BAYES}}(x_{n+1}|x^n) \notin \mathcal{M}$ (strategy outside model).
  - $\mathcal{R}(P_{\text{BAYES}}, x^n) = \frac{k}{2} \log n + O(1)$ (asympt. optimal).
- Plug-in strategy:

$$P_{\text{PLUG-IN}}(x_{n+1} \mid x^n) = P_{\bar{\mu}(x^n)}(x_{n+1}), \quad \bar{\mu} \colon \mathcal{X}^\infty \to \Theta$$

  - $U_{\text{PLUG-IN}}(x_{n+1} \mid x^n) \in \mathcal{M}$ (in-model strategy).
  - ML plug-in strategy ("follow the leader") if $\bar{\mu}(x^n) = \hat{\mu}_n^\circ$:

$$\hat{\mu}_n^\circ = \frac{n_0 x_0 + \sum_{i=1}^n x_i}{n_0 + n} \qquad \text{(smoothed ML estimator)}$$

  - $\mathcal{R}(P_{\text{PLUG-IN}}, , x^n) \geq c \frac{k}{2} \log n + O(1)$, worst case: $c \gg 1$.

Contribution

Bayes strategy:
(strategy outside the model)
- asympt. optimal regret:
$$\frac{k}{2}\log n + O(1)$$
- usually hard to calculate

Plug-in strategy (incl. ML):
(strategy in the model)
- suboptimal:
$$c\frac{k}{2}\log n + O(1)$$
- simple to compute/update

"Follow the Flattened Leader"

A slight modification ("flattening") of the ML plug-in strategy, "almost" in the model, achieving optimal regret.
- achieves performance of Bayes
- retains simplicity of ML

Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

ML strategy prediction:

$$\mathcal{N}(\hat{\mu}_n^\circ, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}_n^\circ, 1 + \frac{1}{n+1}\right)$$

Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

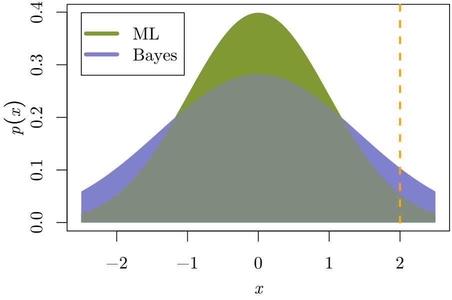
ML strategy prediction:

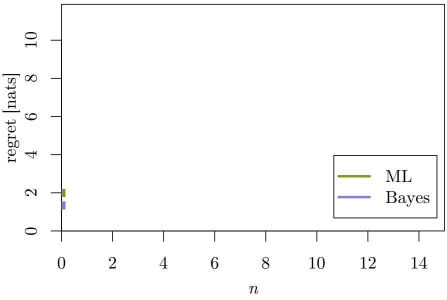
$$\mathcal{N}(\hat{\mu}^\circ_n, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}^\circ_n, 1 + \frac{1}{n+1}\right)$$

- Sequence of outcomes x_n: $2, -2, 2, -2, \ldots$.





Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

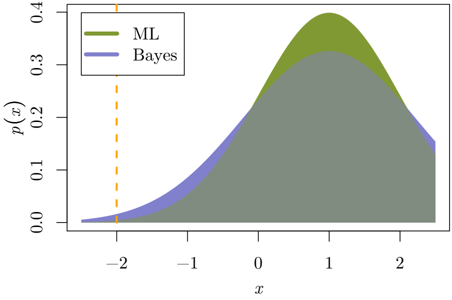
ML strategy prediction:

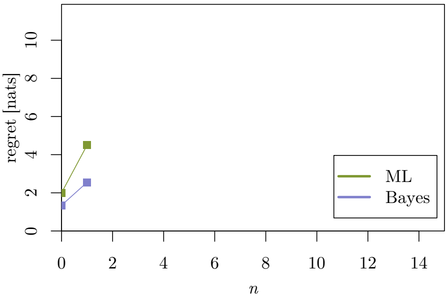
$$\mathcal{N}(\hat{\mu}_n^{\circ}, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}_n^{\circ}, 1 + \frac{1}{n+1}\right)$$

- Sequence of outcomes x_n: $2, -2, 2, -2, \ldots$.





Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$.

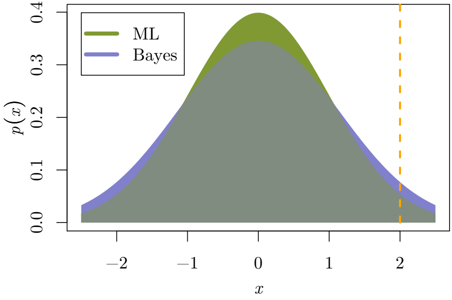
ML strategy prediction:

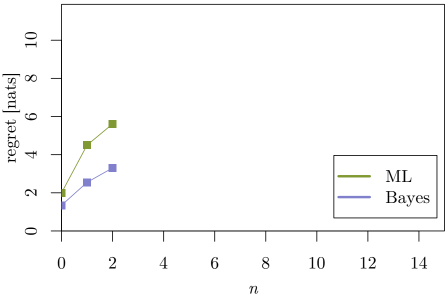
$$\mathcal{N}(\hat{\mu}_n^\circ, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}_n^\circ, 1 + \frac{1}{n+1}\right)$$

- Sequence of outcomes x_n: $2, -2, 2, -2, \ldots$





Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

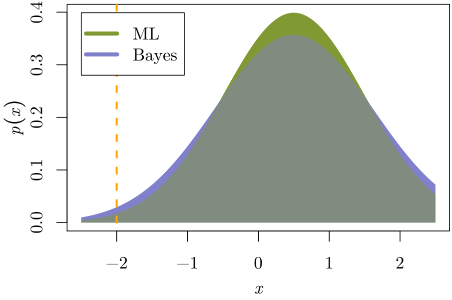
ML strategy prediction:

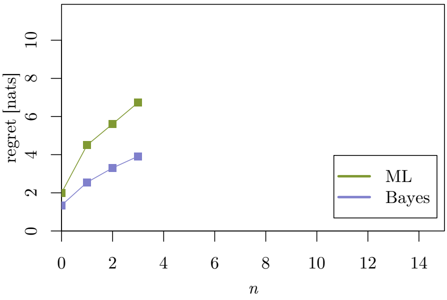
$$\mathcal{N}(\hat{\mu}_n^{\circ}, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}_n^{\circ}, 1 + \frac{1}{n+1}\right)$$

- Sequence of outcomes x_n: $2, -2, 2, -2, \ldots$.





Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

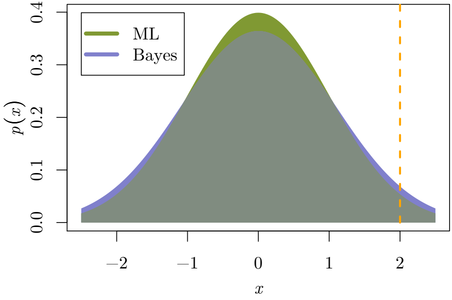
ML strategy prediction:

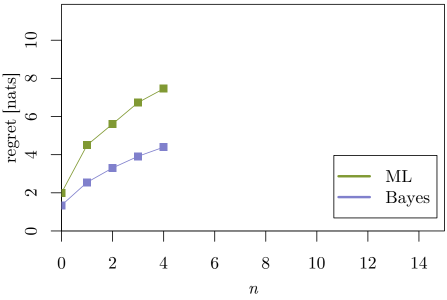
$$\mathcal{N}(\hat{\mu}_n^{\circ}, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}_n^{\circ}, 1 + \frac{1}{n+1}\right)$$

- Sequence of outcomes x_n: $2, -2, 2, -2, \ldots$





Motivating Example: Why Bayes is better than ML?

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

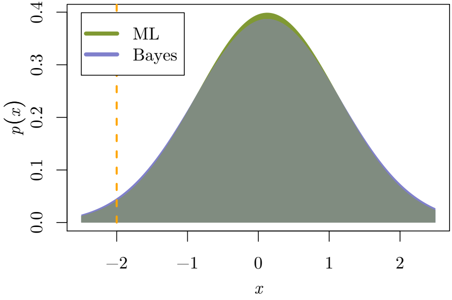
ML strategy prediction:

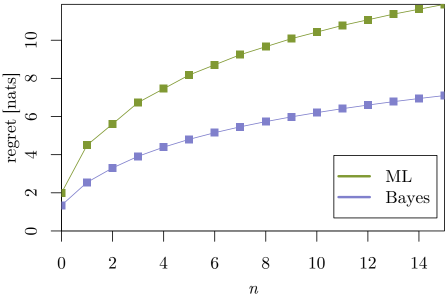
$$\mathcal{N}(\hat{\mu}_n^\circ, 1)$$

Bayes strategy prediction:

$$\mathcal{N}\left(\hat{\mu}_n^\circ, 1 + \frac{1}{n+1}\right)$$

- Sequence of outcomes x_n: $2, -2, 2, -2, \ldots$.





Suboptimal Performance of ML Strategy

Grünwald & de Rooij (2005); Grünwald & Kotłowski (2010)

- $\mathcal{M}$ is a single-parameter exponential family,
- $X_1, X_2, \ldots$ i.i.d. $\sim P^*$, $E_{P^*}[X] = \mu^* \in \Theta$.

$$E_{P^*}[\mathcal{R}(P_{\text{PLUG-IN}}, X^n)] \geq \frac{1}{2} \frac{\mathsf{var}_{P^*} X}{\mathsf{var}_{P_{\mu^*}} X} \log n + O(1),$$

Inferior performance when the variation of data greater than the variance of $P_{\mu^*} \in \mathcal{M}$.

$\Longrightarrow$ Compensate for variability of the data.

Improving ML Strategy

Flattened ML Strategy

$$P_{\text{FML}}(x_{n+1}|x_n) := P_{\hat{\mu}_n^\circ}(x_{n+1}) \frac{n + n_0 + \frac{1}{2}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ)}{n + n_0 + \frac{k}{2}}$$

Improving ML Strategy

compensation for variability of data

for exponential families, $I(\mu) = \mathrm{Cov}_\mu^{-1} X$

Flattened ML Strategy

$$P_{\text{FML}}(x_{n+1}|x_n) := P_{\hat{\mu}_n^\circ}(x_{n+1}) \underbrace{\frac{n + n_0 + \frac{1}{2}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ)}{n + n_0 + \frac{k}{2}}}_{\text{Flattening term: } 1 + O\left(\frac{1}{n}\right)}$$

Main Result: Adversarial Case

Assumptions on outcomes: For all large n:

- sequence of data bounded: $\|x_n\| \leq B$
- sequence of ML estimators $\hat{\mu}_n$ bounded away from $\partial\Theta$.

Then, the flattened ML strategy P_{FML} achieves asymptotically optimal regret, i.e.

$$\mathcal{R}(P_{\mathrm{FML}}, x^n) = \frac{k}{2}\log n + O(1).$$

where the constant under $O(\,\cdot\,)$ does not depend on the outcomes.

Main Result: Stochastic Case

Assumptions on outcomes:

- $X_1, X_2, \ldots$ i.i.d. $\sim P^*$, $E_{P^*}[X] = \mu^* \in \Theta$.
- First four moments of P^* exist.

Then, the flattened ML strategy P_{FML} almost surely achieves asymptotically optimal regret, i.e.

$$\mathcal{R}(P_{\text{FML}}, X^n) = \frac{k}{2} \log n + O(1)$$

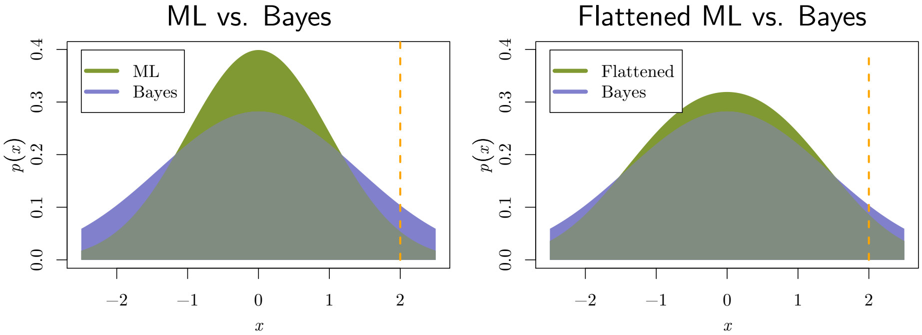
holds with probability one.

Flattened ML vs. ML and Bayes

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.

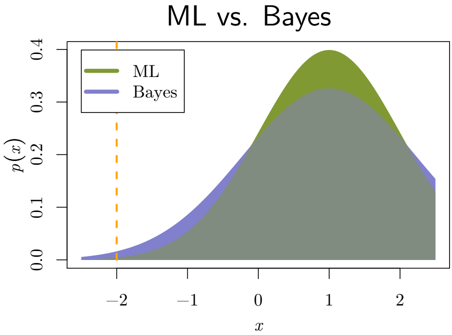
Flattened ML vs. ML and Bayes

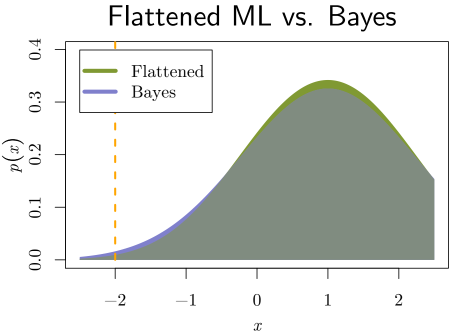
- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.



Flattened ML vs. ML and Bayes

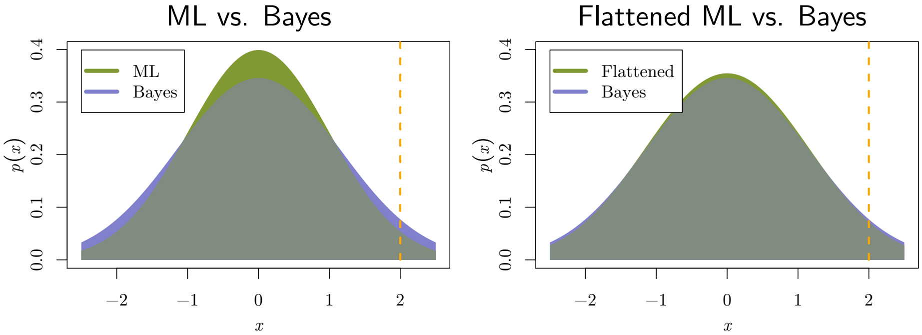
- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.





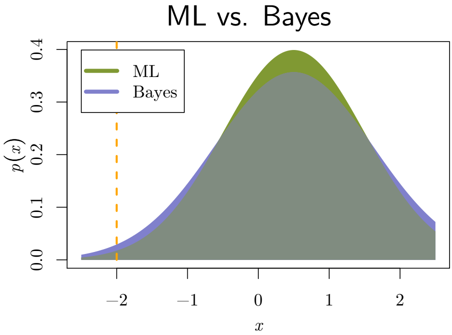
Flattened ML vs. ML and Bayes

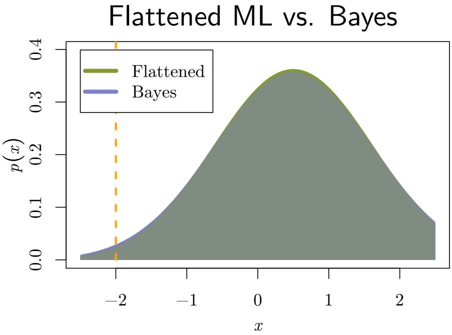
- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.



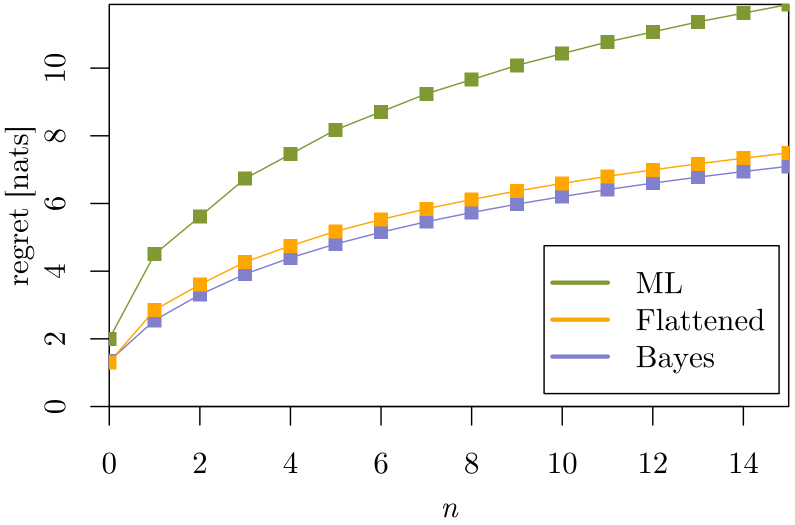
Flattened ML vs. ML and Bayes

- $\mathcal{M} = \{\mathcal{N}(\mu, 1) \colon \mu \in \mathbb{R}\}$.





Flattened ML vs. ML and Bayes



Conclusions

- We proposed a simple "flattening" of the ML distribution for which the optimal asymptotic regret is achieved.
- Flattened ML strategy retains the simplicity of ML strategy, while achieving the performance of Bayes and NML.
- Applications in prediction, coding, model selection.