Motivation: Controlling Complexity

- For many structured prediction tasks, we need very complicated models.
  - We require multiple forms of regularization to prevent overfitting.

- In 340 we saw two ways to reduce complexity of a model:
  - Model averaging (ensemble methods).
  - Regularization (linear models).

- Bayesian methods combine both of these.
  - Average over models, weighted by posterior (which includes regularizer).
  - Allows you to fit extremely-complicated models without overfitting.
Bayesian learning includes:

- Gaussian processes.
- Approximate inference.
- Bayesian nonparametrics.
Why Bayesian Learning?

- Standard L2-regularized logistic regression setup:
  - Given finite dataset containing IID samples.
  - E.g., samples \((x^i, y^i)\) with \(x^i \in \mathbb{R}^d\) and \(y^i \in \{-1, 1\}\).
  - Find “best” \(w\) by minimizing NLL with a regularizer to “prevent overfitting”.

\[
\hat{w} \in \text{argmin}_w \sum_{i=1}^{n} \log p(y^i | x^i, w) + \frac{\lambda}{2} \|w\|^2.
\]

- Predict labels of new example \(\tilde{x}\) using single weights \(\hat{w}\),

\[
\hat{y} = \text{sgn}(\hat{w}^T \tilde{x}).
\]

- But data was random, so weight \(\hat{w}\) is a random variable.
  - This might put our trust in a \(\hat{w}\) where posterior \(p(\hat{w} | X, y)\) is tiny.

- Bayesian approach: “all parameters are nuisance parameters”.
  - Treat \(w\) as random and predict based on rules of probability.
Problems with MAP Estimation

Does MAP make the right decision?

- Consider three hypotheses $\mathcal{H} = \{"lands"", "crashes"", "explodes"\}$ with posteriors:

  \[ p("lands" | D) = 0.4, \quad p("crashes" | D) = 0.3, \quad p("explodes" | D) = 0.3. \]

- The MAP estimate is "plane lands", with posterior probability 0.4.
  - But probability of dying is 0.6.
  - If we want to live, MAP estimate doesn’t give us what we should do.

- Bayesian approach considers all models: says don’t take plane.
- Bayesian decision theory: accounts for costs of different errors.
MAP vs. Bayes

- **MAP** (regularized optimization) approach maximizes over $w$:
  \[
  \hat{w} \in \arg\max_w p(w \mid X, y) \\
  \equiv \arg\max_w p(y \mid X, w)p(w) \quad \text{(Bayes’ rule, $w \perp X$)} \\
  \]

- **Bayesian** approach predicts by integrating over possible $w$:
  \[
  p(\tilde{y} \mid \tilde{x}, X, y) = \int_w p(\tilde{y}, w \mid \tilde{x}, X, y)dw \\
  = \int_w p(\tilde{y} \mid w, \tilde{x}, X, y)p(w \mid \tilde{x}, X, y)dw \quad \text{marginalization rule} \\
  = \int_w p(\tilde{y} \mid w, \tilde{x})p(w \mid X, y)dw \quad \text{product rule} \\
  = \int_w p(\tilde{y} \mid w, \tilde{x}, w)p(w \mid X, y)dw \\
  \tilde{y} \perp X, y \mid \tilde{x}, w
  \]

- Considers all possible $w$, and weights prediction by posterior for $w$. 
Motivation for Bayesian Learning

Motivation for studying Bayesian learning:

1. **Optimal decisions** using rules of probability (and possibly error costs).
2. Gives estimates of variability/confidence.
   - E.g., this gene has a 70% chance of being relevant.
3. Elegant approaches for **model selection** and **model averaging**.
   - E.g., optimize $\lambda$ or optimize grouping of $w$ elements.
4. Easy to **relax IID assumption**.
   - E.g., hierarchical Bayesian models for data from different sources.
5. **Bayesian optimization**: fastest rates for some non-convex problems.
6. Allows models with **unknown/infinite number of parameters**.
   - E.g., number of clusters or number of states in hidden Markov model.

Why isn’t everyone using this?

- Philosophical: Some people don’t like “subjective” prior.
- Computational: Typically leads to nasty integration problems.
Coin Flipping Example: MAP Approach

- MAP vs. Bayesian for a simple coin flipping scenario:
  1. Our likelihood is a Bernoulli,
     \[ p(H \mid \theta) = \theta. \]
  2. Our prior assumes that we are in one of two scenarios:
     - The coin has a 50% chance of being fair ($\theta = 0.5$).
     - The coin has a 50% chance of being rigged ($\theta = 1$).
  3. Our data consists of three consecutive heads: ‘HHH’.

- What is the probability that the next toss is a head?
  - MAP estimate is $\hat{\theta} = 1$, since $p(\theta = 1 \mid HHH) > p(\theta = 0.5 \mid HHH)$.
  - So MAP says the probability is 1.

- But MAP overfits: we believed there was a 50% chance the coin is fair.
Bayesian method needs posterior probability over \( \theta \),

\[
p(\theta = 1 \mid HHH) = \frac{p(HHH \mid \theta = 1)p(\theta = 1)}{p(HHH)} \quad \text{(Bayes rule)}
\]

(marg and prod rule) \[
= \frac{p(HHH \mid \theta = 1)p(\theta = 1)}{p(HHH \mid \theta = 0.5)p(\theta = 0.5) + p(HHH \mid \theta = 1)p(\theta = 1)}
\]

\[
= \frac{(1)(0.5)}{(1/8)(0.5) + (1)(0.5)} = \frac{8}{9},
\]

and similarly we have \( p(\theta = 0.5 \mid HHH) = \frac{1}{9} \).

So given the data, we should believe with probability \( \frac{8}{9} \) that coin is rigged.

There is still a \( \frac{1}{9} \) probability that it is fair that MAP is ignoring.
Coin Flipping Example: Posterior Predictive

- **Posterior predictive** gives probability of head given data and prior,

\[
p(H | HHH) = p(H, \theta = 1 | HHH) + p(H, \theta = 0.5 | HHH)
\]
\[
= p(H | \theta = 1, HHH)p(\theta = 1 | HHH)
\]
\[
+ p(H | \theta = 0.5, HHH)p(\theta = 0.5 | HHH)
\]
\[
= (1)(8/9) + (0.5)(1/9) = 0.94.
\]

- So the correct probability given our assumptions/data is 0.94, and not 1.
  - Though with a different prior we would get a different answer.

- Notice that there was **no optimization** of the parameter \( \theta \):
  - In Bayesian stats we condition on data and integrate over unknowns.

- In Bayesian stats/ML: “all parameters arenuisance parameters”.
Coin Flipping Example: Discussion

Comments on coin flipping example:
- Bayesian prediction uses that HHH could come from fair coin.
- As we see more heads, posterior converges to 1.
  - MLE/MAP/Bayes usually agree as data size increases.
- If we ever see a tail, posterior of $\theta = 1$ becomes 0.

- If the prior is correct, then Bayesian estimate is optimal:
  - Bayesian decision theory gives optimal action incorporating costs.
- If the prior is incorrect, Bayesian estimate may be worse.
  - This is where people get uncomfortable about “subjective” priors.

- But MLE/MAP are also based on “subjective” assumptions.
Bayesian Model Averaging

- In 340 we saw that **model averaging** can improve performance.
  - E.g., random forests average over random trees that overfit.

- But should all models get equal weight?
  - What if we find a **random decision stump** that fits the data perfectly?
    - Should this get the same weight as deep random trees that likely overfit?

- In science, research may be fraudulent or not based on evidence.
  - Should “vaccines cause autism” or “climate change denial” models get equal weight?

- In these cases, naive **averaging may do worse.**
Bayesian Model Averaging

Suppose we have a set of $m$ probabilistic classifiers $w_j$

- Previously our ensemble method gave all models equal weights,

$$p(\tilde{y} \mid \tilde{x}) = \frac{1}{m} p(\tilde{y} \mid \tilde{x}, w_1) + \frac{1}{m} p(\tilde{y} \mid \tilde{x}, w_2) + \cdots + \frac{1}{m} p(\tilde{y} \mid \tilde{x}, w_m).$$

- **Bayesian model averaging** (following rules of probability) weights by posterior,

$$p(\tilde{y} \mid \tilde{x}) = p(w_1 \mid X, y)p(\tilde{y} \mid \tilde{x}, w_1) + p(w_2 \mid X, y)(\tilde{y} \mid \tilde{x}, w_2) + \cdots + p(w_m \mid X, y)p(\tilde{y} \mid \tilde{x}, w_m).$$

So we should weight by probability that $w_j$ is the correct model.

- Equal weights assume all models are equally probable and fit data equally well.
Bayesian Model Averaging

- Weights are posterior, so proportional to likelihood times prior:

\[
p(w_j \mid X, y) \propto p(y \mid X, w_j) p(w_j).
\]

- Likelihood gives more weight to models that predict \(y\) well.
- Prior should give less weight to models that are likely to overfit.

- This is how rules of probability say we should weight models.
  - It's annoying that it requires a “prior” belief over models.
  - But as \(n \to \infty\), all weight goes to “correct” model[s] \(w^*\) as long as \(p(w^*) > 0\).
Bayes for Density Estimation and Generative/Discriminative

- We can use Bayesian approach for **density estimation**:  
  - With data $D$ and parameters $\theta$ we have:  
    1. Likelihood $p(D | \theta)$.  
    2. Prior $p(\theta)$.  
    3. Posterior $p(\theta | D)$.

- We can use Bayesian approach for **supervised learning**:  
  - **Generative** approach (naive Bayes, GDA) are density estimation on $X$ and $y$:  
    1. Likelihood $p(y, X | w)$.  
    2. Prior $p(w)$.  
    3. Posterior $p(w | X, y)$.  
  - **Discriminative** approach (logistic regression, neural nets) just conditions on $X$:  
    1. Likelihood $p(y | X, w)$.  
    2. Prior $p(w)$.  
    3. Posterior $p(w | X, y)$.  

7 Ingredients of Bayesian Inference (MEMORIZE)

1. **Likelihood** $p(y \mid X, w)$.
   - Probability of seeing data given parameters.

2. **Prior** $p(w \mid \lambda)$.
   - Belief that parameters are correct before we’ve seen data.

3. **Posterior** $p(w \mid X, y, \lambda)$.
   - Probability that parameters are correct after we’ve seen data.
   - We won’t use the MAP “point estimate”, we want the whole distribution.

4. **Predictive** $p(\tilde{y} \mid \tilde{x}, w)$.
   - Probability of test label $\tilde{y}$ given parameters $w$ and test features $\tilde{x}$.
     - For example, sigmoid function for logistic regression.
7 Ingredients of Bayesian Inference (MEMORIZE)

5 Posterior predictive $p(\hat{y} | \tilde{x}, X, y, \lambda)$.
- Probability of new data given old, integrating over parameters.
- This tells us which prediction is most likely given data and prior.

6 Marginal likelihood $p(y | X, \lambda)$ (also called “evidence”).
- Probability of seeing data given hyper-parameters (integrating over parameters).
- We’ll use this later for hypothesis testing and setting hyper-parameters.

7 Cost $C(\hat{y} | \tilde{y})$.
- The penalty you pay for predicting $\hat{y}$ when it was really was $\tilde{y}$.
- Leads to Bayesian decision theory: predict to minimize expected cost.
Review: Decision Theory

- Are we equally concerned about “spam” vs. “not spam”.

- Consider a scenario where different predictions have different costs:

<table>
<thead>
<tr>
<th>Predict / True</th>
<th>True “spam”</th>
<th>True “not spam”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict “spam”</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Predict “not spam”</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

- In 340 we discussed predicting $\hat{y}$ given $\hat{w}$ by minimizing expected cost:

$$\mathbb{E}[\text{Cost}(\hat{y} = \text{“spam”})] = p(\tilde{y} = \text{“spam”} | \tilde{x}, \tilde{w})C(\hat{y} = \text{“spam”} | \tilde{y} = \text{“spam”})$$

$$+ p(\tilde{y} = \text{“not spam”} | \tilde{x}, \tilde{w})C(\hat{y} = \text{“spam”} | \tilde{y} = \text{“not spam”}).$$

- Consider a case where $p(\tilde{y} = \text{“spam”} | \tilde{x}, \tilde{w}) > p(\tilde{y} = \text{“not spam”} | \tilde{x}, \tilde{w})$.
  - We might still predict “not spam” if expected cost is lower.
Bayesian Decision Theory

- Bayesian decision theory:
  - Instead of using a MAP estimate $\hat{w}$, we should use posterior predictive,
    
    $$\mathbb{E}[\text{Cost}(\hat{y} = \text{“spam”})] = p(\tilde{y} = \text{“spam”} \mid \tilde{x}, X, y)C(\hat{y} = \text{“spam”} \mid \tilde{y} = \text{“spam”})$$
    $$+ p(\tilde{y} = \text{“not spam”} \mid \tilde{x}, X, y)C(\hat{y} = \text{“spam”} \mid \tilde{y} = \text{“not spam”}).$$

- Minimizing this expected cost is the optimal action.

- Note that there is a lot going on here:
  - Expected cost depends on cost and posterior predictive.
  - Posterior predictive depends on predictive and posterior
  - Posterior depends on likelihood and prior.
Summary

**Bayesian statistics:**
- Condition on the data, integrate (rather than maximize) over posterior.
- “All parameters are nuisance parameters”.

**Bayesian model averaging and decision theory:**
- Model averaging and decision theory based on rules of probability.

**Next time:** learning the prior?