# Gaussians and Bayesian learning

## CPSC 440/550: Advanced Machine Learning

cs.ubc.ca/~dsuth/440/24w2

University of British Columbia, on unceded Musqueam land

2024-25 Winter Term 2 (Jan–Apr 2025)

# So far

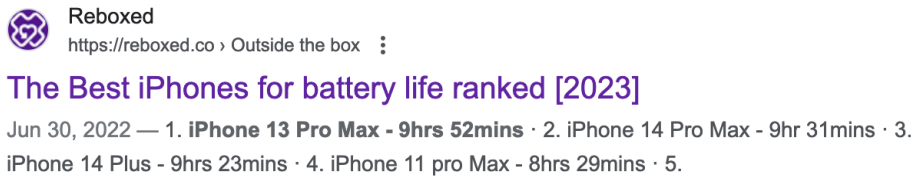
- We've covered binary and categorical random variables
  - Plus a few continuous things to use as priors: beta and Dirichlet
- Use in density estimation
  - Generative model: estimate joint density $p(x, y)$, can use for $p(y \mid x)$
  - Discriminative model: parameterize $p(y \mid x)$ as a function of $x$, do density estimation
  - Bernoulli likelihood for binary classification, categorical (with softmax) for multiclass
- Talked about priors for MAP learning

- Enough to do some really complicated things
- But still missing some important aspects!
- What about when outputs $y$ aren't binary/categorical?

# Motivating problem: phone battery life

- How long until my phone dies?
  - Could model it as "0-30 minutes", "31-60 minutes", "1-2 hours", ...
  - Or "0-1 minutes", "1-2 minutes", "2-3 minutes", ...
  - Probably more sensible to think of it as a continuous quantity

- Usually reviews, ads/reviews give a point estimate:




- But of course the actual time varies
- "If it's at 31% now, what's the probability it'll still have charge in four hours?"

# General problem: continuous density estimation

- Can view the basic version of this as a density estimation of a continuous variable

$$\mathbf{X} = \begin{bmatrix} \texttt{12 hr 37 min 12.3 s} \\ \texttt{17 hr 31 min 54.9 s} \\ \texttt{14 hr 17 min 48.3 s} \\ \texttt{ 9 hr 51 min 20.0 s} \end{bmatrix} \xrightarrow{\text{density estimator}} \begin{array}{l} p(X = \texttt{11 hr 17 min 31.8 s}) = 0.12 \\ p(X = \texttt{13 hr 1 min 18.1 s}) = 1.41 \end{array}$$

- This is a *density*, not a probability!
- For continuous distributions, the probability of getting any *exact* number is zero
- Probability of being in an interval $[a, b]$ is $\int_a^b p(x)\mathrm{d}x$

# Continuous density estimation

- Other applications of continuous density estimation:
  - Modeling sizes (birth weight of babies, size of zucchini grown in this field, ...)
  - Modeling how long it takes to do this step of a manufacturing process
  - Modeling income, maybe age, ...
  - Modeling blood pressure, cholesterol level, ...
  - Modeling grades
  - ...

- Often useful even if it's "really" categorical
  - UBC grades are whole integers between 0 and 100
  - But "83" and "84" are much more similar to each other than "61" or "97"
  - Usually easier to predict "83.8" and round
  - (With enough data, "best" model could handle individual numbers separately)

- Bernoulli/categorical distributions can model basically any binary/categorical data
- This is not true for continuous data: lots of possible shapes!
- We'll start with a simple case: Gaussian/normal distributions

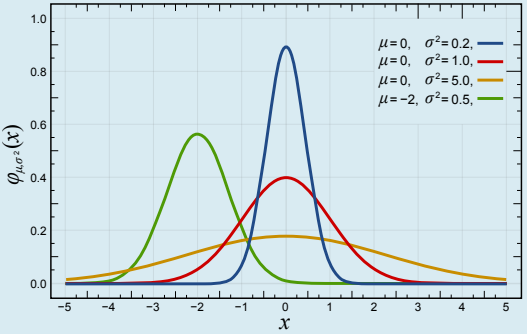
# Univariate Gaussian distribution

review

- A Gaussian random variable, written $X \sim \mathcal{N}(\mu, \sigma^2)$, has density

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

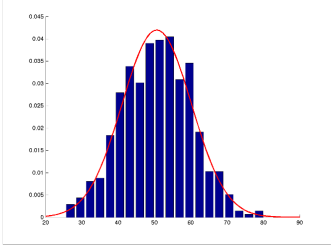
- The mean $\mu = \mathbb{E}[X]$ can be any real number
- The variance $\sigma^2 = \mathrm{Var}(X)$ can be any positive number
  - Sometimes allow $\sigma = 0$; $X$ becomes a point mass, $\Pr(X = \mu) = 1$



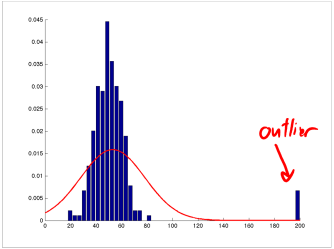
# Why use a Gaussian?

- Your data might actually be Gaussian
  - Great reason to use if it's true! Unfortunately usually not true
- Central limit theorem: many sums of random variables converge to a Gaussian
  - Very often a useful justification for saying e.g. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$ is roughly Gaussian
  - Usually doesn't mean that the data itself is Gaussian
  - Only when your data is approx. the sum of many independent factors
- It's the distribution with maximum entropy for a given mean and variance
  - In some sense, "makes the fewest assumptions" to match given mean and variance
    - We'll return to this soon when we cover exponential families
  - For complicated problems, matching mean and variance isn't enough
- Gaussians make many computations and lots of theory *much* easier
  - Often "good enough to be useful"
  - Very common building block in more advanced methods

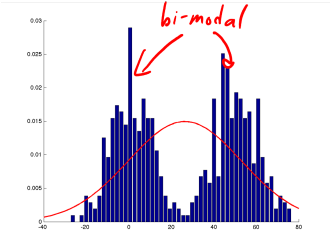
# Why not use a Gaussian?



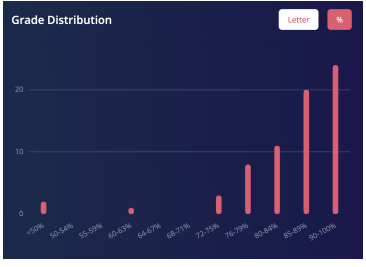
MLE a pretty good fit




sensitive to outliers




can only handle one mode




truncation, asymmetry, outliers

# Gaussian inference

- Decoding the mode: the density $\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$ is maximized if $x = \mu$
- Computing the likelihood of iid data: (now a density, not a probability!)

$$p(\mathbf{X} \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right)$$
$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2\right)$$

- Probability of $X$ in an interval: using the cumulative distribution function (cdf),

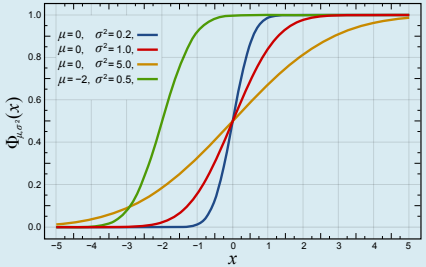
$$\Pr(a \le X \le b \mid \mu, \sigma^2) = \int_a^b p(x \mid \mu, \sigma^2)\mathrm{d}x = \Pr(X \le b \mid \mu, \sigma^2) - \Pr(X \le a \mid \mu, \sigma^2)$$

  - If $a = b$, this is zero (except in the degenerate $\sigma = 0$ case)

# Cumulative distribution functions (cdf)

review

- Often use the cdf $F(t) = \Pr(X \leq t) = \int_{-\infty}^{t} p(x)\mathrm{d}x$
- $F(t)$ is always between $0$ and $1$
- For Gaussians, it's a monotonically increasing function
  - For *any* distribution it's nondecreasing: can't go down, but could stay flat




https://en.wikipedia.org/wiki/Normal_distribution

- For Gaussian, doesn't have an elementary closed form
  - Sometimes written with "error function" $\frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\mathrm{d}t$, but doesn't really help...
  - Get numerically (`scipy.stats.norm.cdf`, `torch.distributions.Normal.cdf`)

# Sampling based on CDFs

- How to sample from a continuous density?
- We want a function that, based on $u \sim \mathrm{Unif}([0,1])$,
  - 50% of the time, returns a sample with $F(x) \leq 0.5$
  - 10% of the time, returns a sample with $0.173 < F(x) \leq 0.273$
  - 1% of the time, returns a sample with $0.8413 \leq F(x) \leq 0.8513$
- That is, we want $F(x)$ to be uniform on $[0,1]$
  - Proof: let $U = F(X)$ for any random variable $X$ with invertible cdf $F$. Then

$$\Pr(U \leq u) = \Pr(F(X) \leq u) = \Pr(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u = \int_0^u 1 \, \mathrm{d}u$$

- If we use $x = F^{-1}(u)$, then $F(x) = F(F^{-1}(u)) = u$ is uniform!
- Inverse transform method for sampling from a 1d continuous density with cdf $F$:
  - Take $u \sim \mathrm{Unif}([0,1])$; return $F^{-1}(u)$
- For Gaussians, no nice form; compute $F^{-1}$ (the "quantile function") numerically
- If can't directly compute the inverse, can do binary search (CDFs are monotonic)
- (Box-Muller transform is more efficient, but Gaussian-specific)

# MLE for univariate Gaussians

- The negative log likelihood (NLL) for $n$ iid samples is

$$-\log p(\mathbf{X} \mid \mu, \sigma^2) = -\log\left(\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2\right)\right)$$

$$= n\log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2 + \text{const}$$

- For any $\sigma$, convex in $\mu$; setting derivative to zero gives sample mean,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

- Plugging in $\hat{\mu}$, setting $\sigma$ derivative to zero gives $\sigma^2$ MLE as the sample variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\hat{\mu})^2$$

  - This step is actually not convex! Need to check that it's still the optimum
  - If all $x^{(i)}$ are the same, get $\hat{\sigma} = 0$; if you require positive $\sigma$, then there's no MLE

# Conjugate prior for the mean

- For fixed variance, conjugate prior for the mean is Gaussian
- If $x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$ are iid, and $\mu \sim \mathcal{N}(m, v)$, then

$$\mu \mid \mathbf{X}, m, v, \sigma^2 \sim \mathcal{N}(\tilde{m}, \tilde{v}), \quad \tilde{m} = \frac{vn}{vn + \sigma^2} \hat{\mu}_{\text{MLE}} + \frac{\sigma^2}{vn + \sigma^2} m, \quad \tilde{v} = \left( \frac{n}{\sigma^2} + \frac{1}{v} \right)^{-1}$$

- Derived by completing the square; see "Gaussians with Conjugate Priors" note
- $\tilde{m}$ is a convex mixture of the prior and the MLE
  - When $n = 0$, it's the prior mean; when $n \to \infty$, it's the MLE
  - MAP is also $\tilde{m}$ (maximizes the posterior density)
- $\tilde{v}$ is half the harmonic mean of $v$ (prior variance) and $\frac{\sigma^2}{n}$ (MLE variance)
  - When $n = 0$, it's the prior variance; when $n \to \infty$, it's zero
- Will return to priors for the variance later

# Supervised learning with Gaussians: generative models

- Can do Gaussian Naïve Bayes with categorical labels: for example,

$$y \sim \mathrm{Cat}(\boldsymbol{\theta}) \qquad x_j \mid y \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

- Everything works just like for binary/categorical data
  - e.g. to fit, do the MLE on each dimension separately for each class

- Can't really do Naïve Bayes with continuous labels!

$$y \sim \mathcal{N}(\mu_y, \sigma_y^2) \qquad x_j \mid y \sim \text{anything}$$

  Only have one sample per $y$ (almost surely); can't really fit the $x$ distributions

- We'll return to Gaussian generative models after multivariate Gaussians

# Supervised learning with Gaussians: discriminative models

- Like before, we can take $y \mid x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ for $\mu_x$, $\sigma_x^2$ functions of $x$
- Negative log likelihood becomes

$$
\begin{aligned}
-\log p(\mathbf{y} \mid \mathbf{X}) &= -\sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left( -\frac{1}{2\sigma_x^2} \left( \mu_x - y^{(i)} \right)^2 \right) \right) \\
&= \sum_{i=1}^{n} \log \sigma_x + \frac{1}{2\sigma_x^2} \left( \mu_x - y^{(i)} \right)^2 + \text{constant}
\end{aligned}
$$

- Linear regression uses $\mu_x = w^\mathsf{T} x$, $\sigma_x = \sigma$ independent of $x$
  - Becomes scaled square loss, plus a constant
- Deep models with square loss also use $\mu_x = f_\theta(x)$, $\sigma_x = \sigma$ independent of $x$
- But can also use $\sigma_x = g_\theta(x)$ to fit!
  - Often share some layers for computation of $\mu_x$ and $\sigma_x$
  - Some challenges with this approach; will discuss a bit more soon

# Linear regression

review

- The usual L2-regularized least squares ("ridge regression") model:

$$y \mid x, w \sim \mathcal{N}\left(w^\mathsf{T} x^{(i)}, \sigma^2\right) \qquad w_j \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{\lambda}\right)$$

$$-\log p(w \mid \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \frac{1}{2\sigma^2}\left(w^\mathsf{T} x^{(i)} - y^{(i)}\right)^2 + \sum_{j=1}^{d} \frac{\lambda}{2} w_j^2 + \text{const}$$

$$= \frac{1}{2\sigma^2}\|\mathbf{X}w - y\|^2 + \frac{\lambda}{2}\|w\|^2$$

- Setting the gradient to zero, if $\lambda > 0$ there's a unique MAP estimate

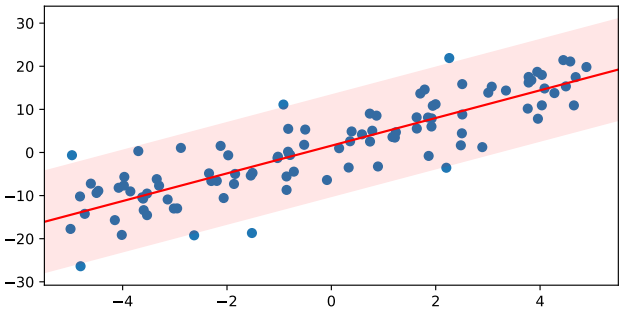
$$\hat{w} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{\lambda}{\sigma^2}\mathbf{I}_d\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\left(\mathbf{X}\mathbf{X}^\mathsf{T} + \frac{\lambda}{\sigma^2}\mathbf{I}_n\right)^{-1} \mathbf{y}$$

and for a new sample $\tilde{x}$, we have $\tilde{y} \mid \tilde{x}, \hat{w} \sim \mathcal{N}(\hat{w}^\mathsf{T}\tilde{x}, \sigma^2)$

# Predictive uncertainty

- MAP estimation allows us to have predictive uncertainty

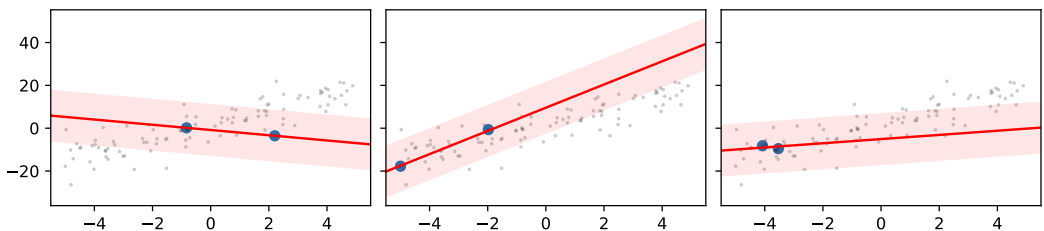
$$y \mid x, w \sim \mathcal{N}\left(w^{\mathsf{T}} x^{(i)}, \sigma^2\right) \qquad w_j \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{\lambda}\right)$$




- Good for modeling "irreducible uncertainty" (also called "aleatoric")
  - ... if $\mathbb{E}[y \mid x]$ is roughly linear, and $y - \mathbb{E}[y \mid x]$ is "Gaussian enough"!
  - Bad if $y \mid x$ is multimodal, unbounded, has heavy tails, ...
  - Also assumes that variance doesn't depend on $x$ ("homoskedastic")

# Predictive uncertainty

- MAP doesn't take into account uncertainty in our model $w$
  - Also called "epistemic uncertainty"
  - $\mathrm{Var}[y \mid x] = \sigma^2$ doesn't depend on $n$
- Do these predictive uncertainties (with $n = 2$) seem reasonable?




- Would like to incorporate uncertainty about $w$ into our predictions

# Bayesian learning

- MAP estimation commits to the single "best" choice of $w$ for its predictions:

$$\hat{w} \in \arg\max_w p(\mathbf{y} \mid \mathbf{X}, w) \qquad \tilde{y} \sim p(\tilde{y} \mid \tilde{x}, \hat{w})$$

- "Fully Bayesian learning" marginalizes out the choice of $w$:

$$\begin{aligned} p(\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y}) &= \int_w p(\tilde{y}, w \mid \tilde{x}, \mathbf{X}, \mathbf{y}) \, \mathrm{d}w \\ &= \int_w p(\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y}, w) \, p(w \mid \tilde{x}, \mathbf{X}, \mathbf{y}) \, \mathrm{d}w \\ &= \int_w p(\tilde{y} \mid \tilde{x}, w) \, p(w \mid \mathbf{X}, \mathbf{y}) \, \mathrm{d}w \end{aligned}$$

- Last line uses standard conditional independence assumptions:
  - $\tilde{y}$ doesn't depend on the training data if we know $w$
  - $\tilde{x}$ doesn't give us any information about $w$
- We weight the predictions of every possible model $w$ by posterior $p(w \mid \mathbf{X}, \mathbf{y})$

# Posterior predictive distribution

- Bayesian learning is based on

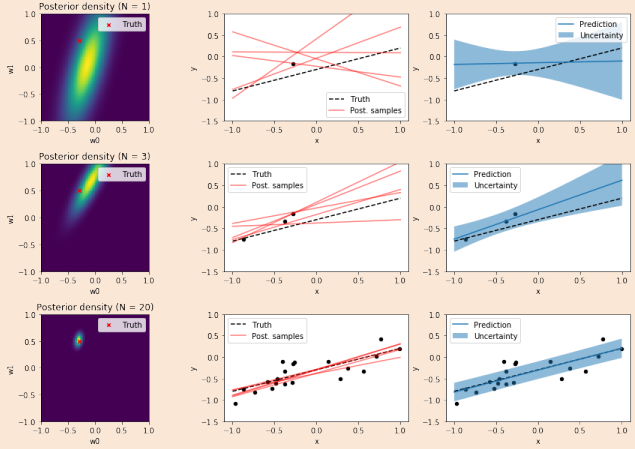
$$p(\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y}) = \int_w p(\tilde{y} \mid \tilde{x}, w)\, p(w \mid \mathbf{X}, \mathbf{y})\, \mathrm{d}w$$

- We call this the posterior predictive distribution
- Could evaluate model quality with $\prod_{i=1}^{n_{test}} p(\tilde{y}^{(i)} \mid \tilde{x}^{(i)}, \mathbf{X}, \mathbf{y})$
- If we have to make a single prediction:
  - The mode $\arg\max_{\tilde{y}} p(\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y})$ would maximize the accuracy, for discrete $y$
  - The mean $\mathbb{E}[\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y}]$ would minimize the expected square loss
  - Might do something else to minimize a different notion of loss

# Bayesian Linear Regression

bonus!

- Bayesian perspective gives us variability in $w$ and predictions:




http://krasserm.github.io/2019/02/23/bayesian-linear-regression

- Will need slightly more mathematical tools to get there; next week!

# Bayesian learning in the Bernoulli-Beta model

- Consider flipping coins with $x \mid \theta \sim \mathrm{Bern}(\theta)$ and prior $\theta \sim \mathrm{Beta}(\alpha, \beta)$
- We showed before that the posterior for $\theta$ is $\theta \mid \mathbf{X} \sim \mathrm{Beta}(\alpha + n_1, \beta + n_0)$
- We can use this to find the posterior predictive, which will be Bernoulli:

$$\begin{aligned} p(\tilde{x} = 1 \mid \mathbf{X}, \alpha, \beta) &= \int_\theta \underbrace{p(\tilde{x} = 1 \mid \theta)}_{\text{prediction}} \underbrace{p(\theta \mid \mathbf{X}, \alpha, \beta)}_{\text{posterior}} \, \mathrm{d}\theta \\ &= \int_\theta \theta \; p_\beta(\theta \mid \alpha + n_1, \beta + n_0) \, \mathrm{d}\theta \\ &= \mathop{\mathbb{E}}_{\theta \sim \mathrm{Beta}(\alpha + n_1, \beta + n_0)}[\theta] = \frac{\alpha + n_1}{\alpha + n_1 + \beta + n_0} = \frac{n_1 + \alpha}{n + \alpha + \beta} \end{aligned}$$

- By comparison: MAP gave the more-confident $\hat{\theta} = \dfrac{n_1 + \alpha - 1}{n + \alpha + \beta - 2}$
- With uniform prior $\alpha = \beta = 1$, MAP is MLE $n_1/n$; Bayesian learning is $\frac{n_1+1}{n+2}$

# Bayesian learning in the Categorical-Dirichlet model

- If $X \mid \boldsymbol{\theta} \sim \mathrm{Cat}(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \mid \boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, we saw before that

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha}) &\propto p(\mathbf{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \propto \theta_1^{n_1} \cdots \theta_k^{n_k} \theta_1^{\alpha_1 - 1} \cdots \theta_1^{n_k - 1} \\ &= \theta_1^{(n_1 + \alpha_1) - 1} \cdots \theta_k^{(n_k + \alpha_k) - 1} \end{aligned}$$

$$\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha} \sim \mathrm{Dir}(\mathbf{n} + \boldsymbol{\alpha}) \qquad \text{where } \mathbf{n} \in \mathbb{R}^d,\ n_j = \sum_{i=1}^{n} \mathbb{1}\left(x^{(i)} = j\right)$$

- MAP: $$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\theta \mid \mathbf{X}) \propto \mathbf{n} + \boldsymbol{\alpha} - 1$$
- Bayesian learning uses the posterior predictive distribution,

$$\begin{aligned} p(x = c \mid \mathbf{X}, \boldsymbol{\alpha}) &= \int_{\boldsymbol{\theta}} p(x = c \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \theta_c\, p(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\theta} \quad = \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim \mathrm{Dir}(\mathbf{n} + \boldsymbol{\alpha})} [\theta_c] \quad \propto \mathbf{n} + \boldsymbol{\alpha} \end{aligned}$$

# Bayesian learning versus MAP

- MAP estimation corresponds to using a regularizer
- Bayesian learning
  - averages over models (like we saw with random forests in 340)
  - weighting each model by its posterior density: its likelihood times a prior (regularizer)
- Can help learn with very complicated models, while controlling overfitting
- One big disadvantage: this integration can be computationally hard!
  - Even for simple cases like our motivating problem of Bayesian linear regression; more next time

# Ingredients of Bayesian inference

1. Likelihood $p(x \mid \theta)$
   - The most important part: model for what the data looks like
2. Prior $p(\theta)$
   - What do we think the parameters might be, before looking at any data?

These imply by the rules of probability:

- Posterior $p(\theta \mid \mathbf{X})$
  - What do we think the parameters might be, after looking at the data?
  - MAP uses $\hat{\theta}$ that maximizes this; Bayesian learning uses whole distribution
- Posterior predictive $p(\tilde{x} \mid \mathbf{X})$
  - What do we think the data distribution looks like, after seeing the training data?
  - Marginalizes over all possible parameters

# Proof of uniformity of CDF value

bonus!

- Let $X$ be any continuous variable with cdf $F(x)$, and define $U = F(X)$
- For any $u \in [0, 1]$,

$$\Pr(U \leq u) = \Pr(F(X) \leq u) = \Pr(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$$

- This is exactly the cdf of a $\mathrm{Unif}([0, 1])$ distribution:

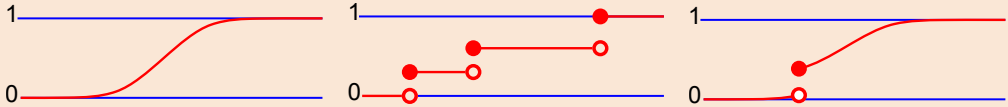
$$\int_0^u 1 \, \mathrm{d}t = u$$

- Equivalent way to see: $p(u) = \frac{\mathrm{d}}{\mathrm{d}u} \Pr(U \leq u) = \frac{\mathrm{d}u}{\mathrm{d}u} = 1$

# Inverse transform sampling for discrete (or mixed) variables

bonus!

- CDFs make sense for discrete, continuous, even mixed variables




https://en.wikipedia.org/wiki/Cumulative_distribution_function

- Discrete values give "jumps" at $\Pr(X \leq x)$, when $\Pr(X = x) > 0$
- CDF is always "right-continuous with left-limits" (RCLL/càdlàg)
- Define quantile function as $Q(u) = \min\{x : u \leq F(x)\}$, which is $F^{-1}$ if $F$ is continuous
- Our "roulette wheel sampling" for categorical distributions is exactly inverse transform sampling: $u \sim \text{Unif}([0, 1])$, return $Q(u)$

# Details on MLE for Gaussians

bonus!

- After plugging in $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$, left with

$$-\log p(\mathbf{X} \mid \mu, \sigma^2) = n \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x^{(i)} - \hat{\mu}\right)^2 + \text{const}$$

$$\propto \log \sigma + \frac{\hat{\sigma}^2}{2\sigma^2} + \text{const} \quad \text{for } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)} - \hat{\mu}\right)^2$$

- Only (finite) stationary points have $\frac{1}{\sigma} - \frac{\hat{\sigma}^2}{\sigma^3} = 0$ so, since $\sigma > 0$, $\sigma^2 = \hat{\sigma}^2$
- Nonconvex ($\frac{\partial^2}{\partial \sigma^2} < 0$ if $\sigma^2 > 3\hat{\sigma}^2$), but enough to check stationary points + limits
  - $\lim_{\sigma \to 0} \left[\log \sigma + \frac{\hat{\sigma}^2}{2\sigma^2}\right] = \infty$ when $\hat{\sigma}^2 > 0$
    - The $\frac{1}{\sigma^2}$ term diverges positively faster than the $\log \sigma$ diverges negatively
    - Write as $\left(\frac{1}{2\sigma^2}\right)\left(\sigma^2 \log \sigma + \hat{\sigma}^2\right)$, have $\lim_{\sigma \to 0} \frac{\log \sigma}{\sigma^{-2}} = \lim_{\sigma \to 0} \frac{\sigma^{-1}}{-2\sigma^{-3}} = \lim_{\sigma \to 0} \frac{-\sigma^2}{2} = 0$ so limit is $\frac{\hat{\sigma}^2}{2\sigma^2} \to \infty$
  - $\lim_{\sigma \to \infty} \left[\log \sigma + \frac{\hat{\sigma}^2}{2\sigma^2}\right] = \infty + 0$