# The Gaussian Distribution from Scratch

Karl Stratos
me@karlstratos.com

Last updated: October, 2023

**Abstract**

The Gaussian distribution has many useful properties. Yet there are few resources that derive these properties from scratch in a concise and comprehensive manner. This technical note is an ongoing effort to develop such a resource. The statements are written as generally as possible, with clean and accessible proofs whenever applicable. Some novel extensions of existing results are provided (e.g., to multivariate forms, non-iid noises). Equipped with the results in this note, we are able to derive complex methods such as diffusion models and sparse Gaussian processes with relative ease, in many cases by simply invoking the Gaussian chain rule and Bayes' rule (which themselves follow beautifully from block matrix operations) instead of calculating any integral.

## Contents

# 1 Definitions

Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}_{\succ 0}$. We assert $\Sigma \succ 0$ (i.e., symmetric and positive-definite) to avoid handling degenerate cases. The **Gaussian distribution** is a mapping $\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to [0, 1]$ defined as

$$\mathcal{N}(\mu, \Sigma)(x) := \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \tag{1}$$

which integrates to 1 over $\mathbb{R}^d$ (Lemma H.4) and is thus a valid probability distribution. The distribution has the moment-generating function (MGF) of $M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$ (Appendix D), which readily shows that $\mu$ is the mean and $\Sigma$ the covariance. If $\Sigma = \operatorname{diag}\left(\sigma_1^2 \dots \sigma_d^2\right)$, (1) is a product distribution of univariate $\mathcal{N}(\mu_i, \sigma_i^2)$. The **standard Gaussian** is the special case with $\mu = 0_d$ and $\Sigma = I_{d\times d}$. The following statements about a random variable $X \in \mathbb{R}^d$ are equivalent (Lemma H.16):

1. $X \sim \mathcal{N}(\mu, \Sigma)$. That is, the probability of $X = x$ is $\mathcal{N}(\mu, \Sigma)(x)$ defined in (1).
2. The MGF of $X$ is $M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$.
3. $X = \mu + \Sigma^{1/2} Z$ where $Z \sim \mathcal{N}(0_d, I_{d\times d})$.
4. $a^\top X \sim \mathcal{N}(a^\top \mu, a^\top \Sigma a)$ for all nonzero $a \in \mathbb{R}^d$.
5. The log probability of $X = x$ is equal to $-\frac{1}{2} x^\top \Sigma^{-1} x + (\Sigma^{-1}\mu)^\top x + C$ where $C \in \mathbb{R}$ is constant in $x$.

If any holds, we say $X \in \mathbb{R}^d$ is **normally distributed** with parameters $(\mu, \Sigma)$. Note that 3 and 4 just reduce general normality to simpler forms of (1) (standard and univariate). These alternative definitions are useful in different contexts, for instance

- 2 shows that a point-mass distribution on $x \in \mathbb{R}^d$ is "normal" with parameters $(x, 0_{d\times d})$, since its MGF is $\mathbf{E}[\exp(t^\top X)] = \exp(t^\top x)$.
- 3 is the popular Gaussian reparameterization trick where we view $X$ as a perturbation of $(\mu, \Sigma)$.
- 4 is handy when showing that $Y$ and $Z$ are jointly normal (Section 3): it is sufficient to show that any scalar projection of $(Y, Z)$ using a nonzero vector is (univariate) normal.
- 5 "completes the square" for you. By putting the log probability of $X$ in this form, we show that $X$ is normal *and* identify its covariance and mean by matching the second- and first-order terms.

# 2 Basic Properties

## 2.1 Shape

(1) implies that the distribution is symmetric: $\mathcal{N}(\mu, \Sigma)(x) = \mathcal{N}(\mu, \Sigma)(-x)$. The gradient and the Hessian matrix of $\mathcal{N}(\mu, \Sigma)$ at $x \in \mathbb{R}^d$ are

$$\begin{aligned}
\left(\nabla \mathcal{N}(\mu, \Sigma)\right)(x) &= -\mathcal{N}(\mu, \Sigma)(x) \times \Sigma^{-1}(x-\mu) \\
\left(\nabla^2 \mathcal{N}(\mu, \Sigma)\right)(x) &= -\mathcal{N}(\mu, \Sigma)(x) \times \left(\Sigma^{-1} - \Sigma^{-1}(x-\mu)(x-\mu)^\top \Sigma^{-1}\right)
\end{aligned}$$

where the Hessian is negative-definite at $x = \mu$ but can be indefinite at other points (Lemma H.32). The distribution is not concave, but it is (strictly) log-concave, thus quasiconcave, with $\mu$ as the unique mode (as well as the mean).

## 2.2 Linear Transformation

A critical property of the Gaussian distribution is that it is closed under linear transformation. Note that definitions 3 and 4 are consistent with this property. For any $A \in \mathbb{R}^{d'\times d}$ and $b \in \mathbb{R}^{d'}$ where $A$ is full-rank with $d' \le d$ (so that $A\Sigma A^\top \succ 0$), $X \sim \mathcal{N}(\mu, \Sigma)$ implies (Lemma D.2):

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top) \tag{2}$$

### 2.3 Log-Likelihood

The Gaussian log-likelihood with a Gaussian random mean is again a Gaussian log-likelihood with a regularization. Pick any $A \in \mathbb{R}^{d' \times d}$ and $\Omega \in \mathbb{R}^{d' \times d'}_{\succ 0}$. For all $y \in \mathbb{R}^{d'}$ (Lemma H.10),

$$\mathop{\mathbf{E}}_{X \sim \mathcal{N}(\mu, \Sigma)} [\log \mathcal{N}(AX, \Omega)(y)] = \log \mathcal{N}(A\mu, \Omega)(y) - \frac{1}{2} \mathrm{tr}\left(\Omega^{-1} A \Sigma A^\top\right) \tag{3}$$

### 2.4 Sample Mean and Covariance

Another characteristic of the Gaussian distribution is that the sample mean and covariance are independent. For any iid $X_1 \dots X_N \sim$ **Unk** with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}_{\succ 0}$, unbiased estimators of the mean and covariance are given by

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i \qquad\qquad \bar{S}^2_N = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)(X_i - \bar{X}_N)^\top$$

It turns out that $\bar{X}_N$ and $\bar{S}^2_N$ are independent iff **Unk** is normal (Geary, 1936). In fact, if **Unk** is normal, then $\bar{X}_N \sim \mathcal{N}(\mu, (1/N)\Sigma)$ and, independently, $(N-1)\bar{S}^2_N \sim \mathcal{W}_d(N-1, \Sigma)$ where $\mathcal{W}_d$ is known as the Wishart distribution (proof).[1] If $d = 1$ and $\Sigma = \sigma^2 > 0$, this implies the better known form $(N-1)/\sigma^2 \bar{S}^2_N \sim \chi^2(N-1)$ where $\chi^2(k)$ is the chi-square distribution with $k$ degrees of freedom.

## 3 Joint Distribution

We say $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are **jointly normally distributed** with parameters $(\mu, \Sigma)$ if the concatenation $(X, Y)$ follows $\mathcal{N}(\mu, \Sigma)$. More explicitly,

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}\right)$$

where $\mu_X \in \mathbb{R}^d$, $\mu_Y \in \mathbb{R}^{d'}$, $\Sigma_X \in \mathbb{R}^{d \times d}_{\succ 0}$, $\Sigma_Y \in \mathbb{R}^{d' \times d'}_{\succ 0}$, $\Sigma_{XY} \in \mathbb{R}^{d \times d'}$, and $\Sigma_{YX} = \Sigma_{XY}^\top$.[2] A subtle fact is that $X$ and $Y$ can be individual normal but not jointly normal (Appendix I), so we must explicitly establish joint normality even for normal variables (e.g., by using 4 or 5). If $X$ and $Y$ are individually normal *and* independent, then they are jointly normal since we can write

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_X & 0_{d \times d'} \\ 0_{d' \times d} & \Sigma_Y \end{bmatrix}\right)$$

If $X, Y$ are jointly normal, uncorrelatedness implies independence (thus they are equivalent).[3] But we must show *joint* normality before claiming independence from uncorrelatedness. For instance, Appendix I gives $X, Y \in \mathbb{R}$ that are individually normal (but not jointly normal) and uncorrelated, but not independent. The following results are useful when inferring independence from uncorrelatedness:

$$\forall A \in \mathbb{R}^{n \times d}, B \in \mathbb{R}^{m \times d}: \quad A\Sigma B^\top = 0_{n \times m} \quad \Leftrightarrow \quad AX \in \mathbb{R}^n \text{ and } BX \in \mathbb{R}^m \text{ are independent} \tag{4}$$

$$\forall A, B \in \mathbb{R}^{d \times d}: \quad A\Sigma B = 0_{d \times d} \quad \Leftrightarrow \quad X^\top AX \in \mathbb{R} \text{ and } X^\top BX \in \mathbb{R} \text{ are independent} \tag{5}$$

where $X \sim \mathcal{N}(\mu, \Sigma)$. Despite their striking similarity, the linear form (4) is simple to prove (Lemma H.11) but the quadratic form (5), known as Craig's theorem (Craig, 1943), is surprisingly difficult and has a long and complicated history (Driscoll and Gundberg Jr, 1986).

[1] Specifically, $\mathcal{W}_d(k, \Sigma)$ is the distribution over $(u_1 \dots u_k)^\top (u_1 \dots u_k) \in \mathbb{R}^{d \times d}$ where $u_1 \dots u_k \in \mathbb{R}^d$ are iid samples from $\mathcal{N}(0_d, \Sigma)$.
[2] We must have $\Sigma_X, \Sigma_Y \succ 0$ since they are main-diagonal blocks of $\Sigma \succ 0$ (Lemma H.9) and $\Sigma_{XY} = \Sigma_{YX}^\top$ since $\Sigma$ is symmetric.
[3] This follows from the form of the conditional distribution (7):

$$\Sigma_{XY} = 0_{d \times d'} \qquad \Rightarrow \qquad \mathcal{N}(\mu, \Sigma)(y|x) = \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})(y) = \mathcal{N}(\mu_Y, \Sigma_Y)(y)$$

### 3.1 Linear Combinations

Let $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{p\times d'}$, and $b \in \mathbb{R}^p$ where $A, B$ are full-rank with $p \le \min(d, d')$. If $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normal with parameters $(\mu, \Sigma)$, we have from (2) that

$$AX + BY + b \sim \mathcal{N}(A\mu_X + B\mu_Y + b, A\Sigma_X A^\top + A\Sigma_{XY}B^\top + B\Sigma_{YX}A^\top + B\Sigma_Y B^\top) \tag{6}$$

In particular, if $X$ and $Y$ are independently normal, then their sum is distributed as

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$$

Note that we need *joint* normality to guarantee the normality of a linear combination. In general a linear combination of normal variables may not be normal (e.g., (118)).

### 3.2 Chain Rule

If $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normal with parameters $(\mu, \Sigma)$, and if $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible, then $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and (Lemma H.12)

$$Y|X = x \sim \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}) \tag{7}$$

(7) can be expressed more simply in terms of the blocks of the precision matrix $\Lambda = \Sigma^{-1}$. In this case (Lemma H.13),

$$Y|X = x \sim \mathcal{N}\left(\mu_Y - \Lambda_Y^{-1}\Lambda_{YX}(x - \mu_X),\ \Lambda_Y^{-1}\right) \tag{8}$$

### 3.3 Bayes' Rule

If $X \sim \mathcal{N}(\mu, \Sigma_X)$ ("Gaussian prior") and $Y|X = x \sim \mathcal{N}(Ax+b, \Sigma_Y)$ ("linear-Gaussian likelihood") where $A \in \mathbb{R}^{d'\times d}$ and $b \in \mathbb{R}^{d'}$, then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_X A^\top \\ A\Sigma_X & \Sigma_Y + A\Sigma_X A^\top \end{bmatrix}\right)$$

along with the marginal and posterior distributions

$$Y \sim \mathcal{N}(A\mu + b, \Sigma_Y + A\Sigma_X A^\top) \tag{9}$$

$$X|Y = y \sim \mathcal{N}(\Lambda_X^{-1}(\Sigma_X^{-1}\mu + A^\top\Sigma_Y^{-1}(y - b)),\ \Lambda_X^{-1}) \tag{10}$$

where $\Lambda_X = \Sigma_X^{-1} + A^\top\Sigma_Y^{-1}A$ (Lemma H.14). In particular, the Gaussian prior is conjugate for the linear-Gaussian likelihood.

## 4 Entropy

Let $\mu' \in \mathbb{R}^d$ and $\Sigma' \in \mathbb{R}^{d\times d}_{\succ 0}$ be parameters of an additional Gaussian distribution over $\mathbb{R}^d$. Then the cross entropy between $\mathcal{N}(\mu', \Sigma')$ and $\mathcal{N}(\mu, \Sigma)$ is (Lemma H.6):

$$H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2}(\mu' - \mu)^\top\Sigma^{-1}(\mu' - \mu) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma'\right) + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \tag{11}$$

This is sufficient to derive entropy and KL divergence:

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma)\right) \tag{12}$$

$$\mathrm{KL}(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2}(\mu' - \mu)^\top\Sigma^{-1}(\mu' - \mu) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma' - I_{d\times d}\right) + \frac{1}{2}\log\left(\frac{\det(\Sigma)}{\det(\Sigma')}\right) \tag{13}$$

Notably, $\mathcal{N}(\mu, \Sigma)$ has the largest entropy among all distributions over $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma$ (Theorem C.1). This is mainly because it standardizes $x$ inside the exponential function.

### 4.1 Mutual Information

Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be jointly normal with parameters $(\mu, \Sigma)$. If $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible, then conditional entropy and mutual information are (Lemma H.15):

$$H(Y|X=x) = \frac{1}{2}\log\left((2\pi e)^{d'}\det(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right) \tag{14}$$

$$I(X,Y) = \frac{1}{2}\log\left(\frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)}\right) \tag{15}$$

where $x \in \mathbb{R}^d$ is arbitrary (so $H(Y|X) = H(Y|X=x)$). Note that $I(X,Y)$ is infinite if $Y = X$. By the noisy-channel coding theorem, mutual information is the capacity (highest information rate that can be achieved nearly error-free) of a communication channel between $X$ and $Y$. Below we give some well-known models with controllable mutual information.

#### 4.1.1 Additive white Gaussian noise channel

Let $X \sim \mathcal{N}(0, \sigma^2)$ and $Z \sim \mathcal{N}(0, \nu^2)$ independently, and define $Y = X + Z$. $X$ and $Y$ are jointly normal because $a_1X + a_2Y = (a_1 + a_2)X + a_2Z$ is a sum of independently normal variables and thus normal for all nonzero $a = (a_1, a_2)$ (definition 4). Since $\mathrm{Var}(Y) = \sigma^2 + \nu^2$ and $\mathrm{Cov}(X,Y) = \sigma^2$,

$$\begin{bmatrix}X\\Y\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2 & \sigma^2\\ \sigma^2 & \sigma^2+\nu^2\end{bmatrix}\right) \qquad \Rightarrow \qquad I(X,Y) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{\nu^2}\right)$$

Thus $I(X, X+Z)$ grows logarithmically in signal-to-noise ratio $\frac{\sigma^2}{\nu^2}$.

#### 4.1.2 Correlated standard normal channel

Let $X, Y \in \mathbb{R}$ be jointly standard normal with correlation $\rho < 1$. One way to construct them is to let $X, Z \overset{iid}{\sim} \mathcal{N}(0,1)$ and set $Y = \rho X + \sqrt{1-\rho^2}Z$. Then

$$\begin{bmatrix}X\\Y\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right) \qquad \Rightarrow \qquad I(X,Y) = -\frac{1}{2}\log\left(1-\rho^2\right)$$

By taking the correlation $\rho \to 1$ we can arbitrarily increase $I(X,Y)$.

## 5 Central Limit Theorem

Let $\mathbf{Unk}(\mu, \sigma^2)$ denote an unknown distribution over $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2 > 0$. It is often of interest to consider the sample average $\bar{X}_N$ defined as

$$X_1 \dots X_N \overset{\text{iid}}{\sim} \mathbf{Unk}(\mu, \sigma^2) \qquad\qquad \bar{X}_N := \frac{1}{N}\sum_{i=1}^{N} X_i$$

The average is itself random: every time we draw $N$ iid samples from $\mathbf{Unk}(\mu, \sigma^2)$, we draw a single sample of $\bar{X}_N$. We can easily verify that $\mathbf{E}[\bar{X}_N] = \mu$ and $\mathrm{Var}\left(\bar{X}_N\right) = \frac{\sigma^2}{N}$, which states that $\bar{X}_N$ concentrates around $\mu$ as $N \to \infty$ (this is called the "law of large numbers"). But what is the *distribution* of $\bar{X}_N$? The **central limit theorem** (CLT) states that $\bar{X}_N$ is asymptotically normal. More precisely, as $N \to \infty$ we have

$$\sqrt{N}\left(\bar{X}_N - \mu\right) \overset{\text{approx.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{16}$$

or, using the closure under linear transformation,

$$\bar{X}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right) \tag{17}$$

which is consistent with but not implied by the law of large numbers. CLT allows us to make probabilistic statements about sample averages regardless of the underlying distribution. For instance, if $X_1 \dots X_N$ are arbitrary iid samples

with mean 42 and variance 7, then approximately $\bar{X}_N \sim \mathcal{N}(42, \frac{7}{N})$ so that we can calculate quantities like $\Pr(\bar{X}_N \leq 50)$ (e.g., by consulting a standard normal table).

A proof of CLT shows that the KL divergence between the distribution of $\sqrt{N}\left(\bar{X}_N - \mu\right)$ and $\mathcal{N}(0, \sigma^2)$ goes to zero as $N \to \infty$. It is nontrivial: we refer to Marsh (2013) for details. CLT generalizes naturally to multivariate. If $\mathbf{Unk}(\mu, \Sigma)$ is an unknown distribution over $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma \succ 0$, then the average $\bar{X}_N$ of samples $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathbf{Unk}(\mu, \Sigma)$ satisfies as $N \to \infty$:

$$\sqrt{N}\left(\bar{X}_N - \mu\right) \overset{\text{approx.}}{\sim} \mathcal{N}(0_d, \Sigma) \tag{18}$$

$$\bar{X}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{1}{N}\Sigma\right) \tag{19}$$

# 6 Exponential Family

An exponential family is any set of distributions over $\mathbb{R}^d$ that can be expressed as

$$q_{h,\tau,\theta}(x) = \underbrace{h(x)}_{\text{base measure } (>0)} \exp\left(\underbrace{\theta^\top}_{\text{natural parameter } (\mathbb{R}^m)} \underbrace{\tau(x)}_{\text{sufficient statistic } (\mathbb{R}^m)} - \underbrace{A_{h,\tau}(\theta)}_{\text{log-partition function}}\right) \tag{20}$$

where the log-partition function has the important property of generating cumulants of the sufficient statistic when differentiated (e.g., $\nabla A_{h,\tau}(\theta)$ is the mean of $\tau(x)$ where $x \sim q_{h,\tau,\theta}$). The set of Gaussian distributions is an exponential family (Appendix E), with one parameterization (Lemma H.23)

$$\mathcal{N}(\mu, \Sigma)(x) = \underbrace{\frac{1}{(\sqrt{2\pi})^d}}_{\text{base measure}} \exp\left(\underbrace{\begin{bmatrix}\Sigma^{-1}\mu \\ -\frac{1}{2}\text{vec}\left(\Sigma^{-1}\right)\end{bmatrix}^\top}_{\text{natural parameter } (\mathbb{R}^{d(d+1)})} \underbrace{\begin{bmatrix}x \\ \text{vec}\left(xx^\top\right)\end{bmatrix}}_{\text{sufficient statistic } (\mathbb{R}^{d(d+1)})} \underbrace{- \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log\left(\det(\Sigma)\right)\right)}_{\text{log-partition function}}\right) \tag{21}$$

where $\text{vec}(A) \in \mathbb{R}^{d^2}$ vectorizes matrix $A \in \mathbb{R}^{d\times d}$. Thus it inherits the usual properties of an exponential family such as the concavity of the likelihood function and the availability of conjugate priors.

## 6.1 Exponential Tilting

Any "base" distribution $p(x)$ can be used as the (normalized) base measure in (20) and, using the identity sufficient statistic $\tau(x) = x$, generates a new exponential family as $g_t(x) \propto e^{t^\top x}p(x)$ indexed by natural parameter $t \in \mathbb{R}^d$. This technique is called exponential tilting. A useful fact is that the Gaussian distributions are closed under exponential tilting (Lemma H.26):

$$\Pr(X_t = x) \propto e^{t^\top x} \times \mathcal{N}(\mu, \Sigma)(x) \qquad \Rightarrow \qquad X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma) \tag{22}$$

### 6.1.1 Aside: Tweedie's formula

(22) can be used to derive a score-based Bayesian estimator called Tweedie's formula, which is a point estimator implied from the following posterior for $\mu \sim g$ and $x|\mu \sim \mathcal{N}(\mu, \Sigma)$ (Lemma E.2)

$$\mu|x \sim \mathbf{Unk}\Big(\underbrace{x + \Sigma\nabla l(x)}_{\text{Tweedie's formula}}, \Sigma(I_{d\times d} + \nabla^2 l(x)\Sigma)\Big)$$

where $l(x) = \log m(x)$ is the log-marginal. It is typically motivated as correcting for "selection bias" (Efron, 2011). Suppose we observe $N$ samples $x_i \sim \mathcal{N}(\mu_i, \sigma^2)$ where the mean itself is drawn from some unknown prior $\mu_i \sim g$ every time. Consider the problem of estimating the mean of $x_{\max} = \max_{i=1}^N x_i$. Maximum-likelihood estimation $\hat{\mu}_{\text{MLE}} = x_{\max}$ almost certainly overestimates the true mean for large $N$. Instead, we can consider the bias-corrected estimator $\hat{\mu} = x_{\max} + \sigma^2 \frac{\partial}{\partial x}(\log m(x))|_{x=x_{\max}}$ where $m(x) = \int_{\mu\in\mathbb{R}} g(\mu)\mathcal{N}(\mu, \sigma^2)(x)d\mu$ is the marginal distribution. Intuitively, we will have $\frac{\partial}{\partial x}(\log m(x))|_{x=x_{\max}} < 0$ because $x_{\max}$ is too large given the knowledge of a shared prior.

# 7 Sub-Gaussian Distributions

A random scalar $S \in \mathbb{R}$ with $\mathbf{E}[S] = 0$ is **sub-Gaussian with variance factor** $\sigma^2$, denoted by $S \sim \mathcal{G}(\sigma^2)$, if

$$\psi_S(t) \leq \psi_{Z \sim \mathcal{N}(0,\sigma^2)}(t) = \frac{\sigma^2 t^2}{2} \tag{23}$$

for all $t \in \mathbb{R}$. It is stable in the following sense:

1. $\mathrm{Var}(S) \leq \sigma^2$ (Lemma H.29).
2. $-S \sim \mathcal{G}(\sigma^2)$. This can be seen by noting that $\psi_{-S}(t) = \psi_S(-t)$.
3. $\Pr(S \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$ for all $\epsilon \geq 0$. Use Chernoff's inequality (H.19) with Lemma H.30 and (90).
4. If $S_1 \dots S_N$ are independent with $S_i \sim \mathcal{G}(\sigma_i^2)$, then $\sum_{i=1}^N S_i \sim \mathcal{G}(\sum_{i=1}^N \sigma_i^2)$.

Combining these properties, we have (Lemma H.31)

$$S_i \sim \mathcal{G}(\sigma_i^2) \text{ independently} \quad \Rightarrow \quad \Pr\left(\left|\frac{1}{N}\sum_{i=1}^N S_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{N^2\epsilon^2}{2\left(\sum_{i=1}^N \sigma_i^2\right)}\right) \tag{24}$$

An important class of sub-Gaussian variables is bounded scalars: if $X \in [a,b]$ then $X - \mathbf{E}[X] \sim \mathcal{G}(\frac{(b-a)^2}{4})$ (Hoeffding's lemma, H.27). This yields the following popular tail inequality.

**Corollary 7.1** (Hoeffding's inequality)**.** If $X_1 \dots X_N \in [a,b]$ are iid with mean $\mu = \mathbf{E}[X_i] \in \mathbb{R}$,

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right) \tag{25}$$

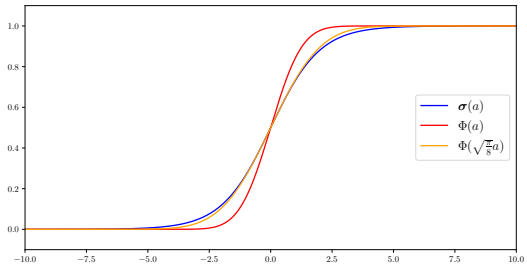
*Proof.* By Hoeffding's lemma, $X_i - \mu \sim \mathcal{G}(\frac{(b-a)^2}{4})$. We get the statement by plugging $\sigma_i^2 = \frac{(b-a)^2}{4}$ in (24). □

# 8 Cumulative Distribution Function

The cumulative distribution function (CDF) $\Phi : (-\infty, \infty) \to (0,1)$ of the standard normal distribution is[4]

$$\Phi(a) := \Pr(X \leq a) = \int_{-\infty}^{a} \mathcal{N}(0,1)(x)dx$$

where $\Phi(0) = \frac{1}{2}$ (by symmetry) and $\Phi'(a) = \mathcal{N}(0,1)(a)$ (by the fundamental theorem of calculus). One use of $\Phi$ is approximating $\boldsymbol{\sigma}(a) := (1 + e^{-a})^{-1}$. We find $\lambda$ so that the slope of $\Phi(\lambda a)$ is the same as that of $\boldsymbol{\sigma}(a)$ at 0. This yields (Lemma H.33)




$$\Phi\left(\sqrt{\frac{\pi}{8}}a\right) \approx \boldsymbol{\sigma}(a) \tag{26}$$

The quality of the approximation is visually apparent in the figure. Another useful property of $\Phi$ is that it is closed under a Gaussian expectation. For any $\lambda, \beta \in \mathbb{R}$ (Lemma H.34):

$$\mathop{\mathbf{E}}_{X \sim \mathcal{N}(\mu,\sigma^2)}\left[\Phi\left(\lambda X + \beta\right)\right] = \Phi\left(\frac{\lambda\mu + \beta}{\sqrt{1 + \lambda^2\sigma^2}}\right) \tag{27}$$

(26) and (27) can be used together to derive an approximate closure of sigmoid under a Gaussian expectation:

$$\mathop{\mathbf{E}}_{X \sim \mathcal{N}(\mu,\sigma^2)}\left[\boldsymbol{\sigma}\left(X\right)\right] \approx \boldsymbol{\sigma}\left(\left(1 + \frac{\pi\sigma^2}{8}\right)^{-1/2}\mu\right) \tag{28}$$

[4]For any distribution $f$ over $X \in \mathbb{R}^d$ the associated CDF is $F(a) := \Pr(X_1 \leq a_1 \wedge \dots \wedge X_d \leq a_d)$. We focus on the one-dimensional standard normal because it is the most useful.

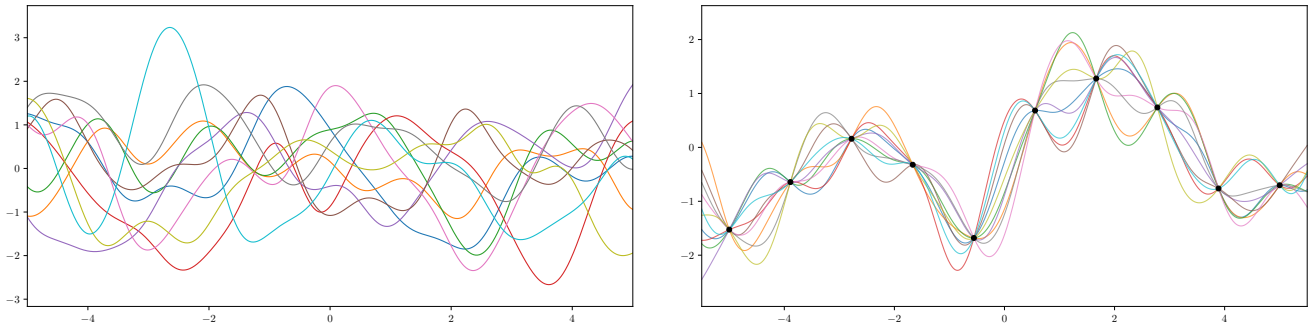


Figure 1: (Left) Ten draws of (29) under the RBF kernel where $x \in \mathbb{R}^{1000}$ is uniformly spaced points in $[-5, 5]$ (representing $\mathcal{X} = \mathbb{R}$). (Right) Ten draws of (30) where $x \in \mathbb{R}^{10}$ is random points, $f \in \mathbb{R}^{10}$ is their scores sampled from (29), and $x_{\text{test}} \in \mathbb{R}^{1000}$ is again uniformly spaced points.

# 9 Gaussian Processes

A **Gaussian process (GP)** is a generalization of joint normality to infinitely many random variables. It assumes a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For concreteness, we will use the RBF kernel $k(u, v) = \exp(-\frac{1}{2\sigma^2} ||u - v||^2)$ over $\mathcal{X} = \mathbb{R}^d$ with the bandwidth hyperparameter $\sigma > 0$. Given a set of $N$ inputs $x \in \mathcal{X}^N$ (no duplicates), a GP defines a conditional distribution over their "scores" $f \in \mathbb{R}^N$ by

$$f \sim \mathcal{N}(0_N, k(x)) \tag{29}$$

where $k(x) \in \mathbb{R}^{N \times N}_{\succ 0}$ denotes the Gram matrix. By defining a marginal distribution over the scores of any finite subset of $\mathcal{X}$, a GP implicitly defines a distribution over *functions* $f : \mathcal{X} \to \mathbb{R}$ (imagine taking $N \to \infty$). The choice of the kernel dictates the function class. Under the RBF kernel, $\text{Var}(f_i) = 1$ (since $k(u, u) = 1$). Thus $\text{cor}(f_i, f_j) = 1$ for infinitesimally close $x_i$ and $x_j$, with the degenerate marginal distribution

$$\begin{bmatrix} f_i \\ f_j \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right)$$

which assigns all the probability mass uniformly on $(f_i, f_j) \in \mathbb{R}^2$ satisfying $f_i = f_j$. (On the other hand, $f_i$ and $f_j$ are independent for infinitely far away $x_i$ and $x_j$.) Hence the sampled functions are *smooth*. See the left plot in Figure 1 for an illustration.

## 9.1 The Predictive Model

Let $x \in \mathcal{N}^N$ and $x_{\text{test}} \in \mathcal{X}^M$ denote training and test inputs. We introduce the latents $f \in \mathbb{R}^N$ and $f_{\text{test}} \in \mathbb{R}^M$ distributed as a GP on $(x, x_{\text{test}}) \in \mathcal{X}^{N+M}$:

$$p_{FF_{\text{test}}}(f, f_{\text{test}}) = \mathcal{N}\left(0_{N+M}, \begin{bmatrix} k(x) & k(x, x_{\text{test}}) \\ k(x_{\text{test}}, x) & k(x_{\text{test}}) \end{bmatrix}\right) \begin{bmatrix} f \\ f_{\text{test}} \end{bmatrix}$$

(we will always omit the implicit conditioning on the inputs). We can easily compute the relations between the latents because of joint normality, for instance by (7):

$$p_{F_{\text{test}}|F}(f_{\text{test}}|f) = \mathcal{N}\left(k(x_{\text{test}}, x)k(x)^{-1}f,\ k(x_{\text{test}}) - k(x_{\text{test}}, x)k(x)^{-1}k(x, x_{\text{test}})\right)(f_{\text{test}}) \tag{30}$$

Note that if $x_{\text{test}} = x$, then (30) reduces to $\mathcal{N}(f, 0_{N \times N})$. See the right plot in Figure 1 for an illustration. Let $r \in \mathcal{R}^N$ denote the training labels where $\mathcal{R}$ denote the label space, and let $p_{R|F}(\cdot|f)$ be the likelihood over the training labels given their scores. For example, we may define

$$p_{R|F}(r|f) = \begin{cases} \mathcal{N}(f, \Sigma)(r) & \text{for regression } (\mathcal{R} = \mathbb{R}) \\ \prod_{i=1}^N \boldsymbol{\sigma}((2r_i - 1)f_i) & \text{for binary classification } (\mathcal{R} = \{0, 1\}) \end{cases} \tag{31}$$

More specifically, the regression labels are a Gaussian perturbation of the training scores: $r = f + \epsilon$ where $\epsilon \sim \mathcal{N}(0_N, \Sigma)$; the classification labels are sampled as $r \sim \text{Ber}(\boldsymbol{\sigma}(f))$ where $\boldsymbol{\sigma} : \mathbb{R} \to (0,1)$ is the sigmoid function. The predictive model under a GP is

$$p_{FF_{\text{test}}R}(f, f_{\text{test}}, r) = p_{FF_{\text{test}}}(f, f_{\text{test}}) \times p_{R|F}(r|f)$$

The (very Bayesian) goal is to estimate the "predictive posterior" $p_{F_{\text{test}}|R}(\cdot|r)$. Under the model, it is given by

$$p_{F_{\text{test}}|R}(f_{\text{test}}|r) = \int_{f \in \mathbb{R}^N} p_{F|R}(f|r) \times p_{F_{\text{test}}|F}(f_{\text{test}}|f)\, df \tag{32}$$

Since $p_{F_{\text{test}}|F}(\cdot|f)$ is Gaussian and known (30), if we also have a Gaussian form of the *posterior* $p_{F|R}(\cdot|r) = \mathcal{N}(\mu^{\text{p}}, \Sigma^{\text{p}})$ where $\mu^{\text{p}} \in \mathbb{R}^m$ and $\Sigma^{\text{p}} \in \mathbb{R}^{m \times m}$ are functions of the training data, then (32) is given by Bayes' rule (9) as $p_{F_{\text{test}}|R}(\cdot|r) = \mathcal{N}(\mu^{\text{pp}}, \Sigma^{\text{pp}})$ where

$$\mu^{\text{pp}} = k(x_{\text{test}}, x)k(x)^{-1}\mu^{\text{p}} \tag{33}$$

$$\Sigma^{\text{pp}} = k(x_{\text{test}}) - k(x_{\text{test}}, x)(k(x)^{-1} - k(x)^{-1}\Sigma^{\text{p}}k(x)^{-1})k(x, x_{\text{test}}) \tag{34}$$

We may then infer the test labels in a task-specific manner. In regression, we may simply return the mean $\mu^{\text{pp}} \in \mathbb{R}^M$ (33). In classification, we compute the expected probabilities for each of the test points $j = 1 \dots M$

$$\Pr\left(j\text{-th test point has label 1}\right) = \mathop{\mathbf{E}}_{f_{\text{test},j} \sim \mathcal{N}(\mu^{\text{pp}}_j, \Sigma^{\text{pp}}_{j,j})} \left[\boldsymbol{\sigma}(f_{\text{test},j})\right] \approx \boldsymbol{\sigma}\left(\left(1 + \frac{\pi \Sigma^{\text{pp}}_{j,j}}{8}\right)^{-1/2} \mu^{\text{pp}}_j\right) \tag{35}$$

(the approximation uses (28)), whereby we return label 1 if (35) is greater than $\frac{1}{2}$ and 0 otherwise. Since $\mu^{\text{p}}$ and $\Sigma^{\text{p}}$ completely determine the solution, we only need to specify the Gaussian posterior (or a Gaussian approximation of the posterior) to make predictions.

### 9.1.1 The regression posterior

In regression, the prior $p_F(\cdot) = \mathcal{N}(0_N, k(x))$ and the likelihood $p_{R|F}(\cdot|f) = \mathcal{N}(f, \Sigma)$ are both Gaussians. Thus the marginal is Gaussian $p_R(\cdot) = \mathcal{N}(0_N, k(x) + \Sigma)$ by (9). The marginal log-likelihood (MLL) $\log p_R(r)$ may be used for tuning kernel hyperparameters. The posterior is also Gaussian $p_{F|R}(\cdot|r) = \mathcal{N}(\mu^{\text{p}}, \Sigma^{\text{p}})$ with the mean and covariance specified by (10):

$$\mu^{\text{p}} = (k(x)^{-1} + \Sigma^{-1})^{-1}\Sigma^{-1}r \qquad \Leftrightarrow \qquad \mu^{\text{p}} = k(x)(k(x) + \Sigma)^{-1}r \tag{36}$$

$$\Sigma^{\text{p}} = (k(x)^{-1} + \Sigma^{-1})^{-1} \qquad\qquad\quad \Sigma^{\text{p}} = k(x) - k(x)(\Sigma + k(x))^{-1}k(x) \tag{37}$$

where the alternative forms are given by the Welling identity (63) and the Woodbury identity (65). Plugging these in (33) and (34), we have the predictive posterior

$$p_{F_{\text{test}}|R}(f_{\text{test}}|r) = \mathcal{N}\left(k(x_{\text{test}}, x)(k(x) + \Sigma)^{-1}r,\ k(x_{\text{test}}) - k(x_{\text{test}}, x)(k(x) + \Sigma)^{-1}k(x, x_{\text{test}})\right)(f_{\text{test}}) \tag{38}$$

which generalizes Bayesian linear regression (106) and reduces to (30) in the "noiseless" setting (i.e., $\Sigma = 0_{N \times N}$). Another way to derive (38) is to apply the Gaussian chain rule (7) on the following observation:

$$\begin{bmatrix} r \\ f_{\text{test}} \end{bmatrix} \sim \mathcal{N}\left(0_{N+M}, \begin{bmatrix} k(x) + \Sigma & k(x, x_{\text{test}}) \\ k(x_{\text{test}}, x) & k(x_{\text{test}}) \end{bmatrix}\right) \tag{39}$$

### 9.1.2 The classification posterior

In classification, the likelihood $p_{R|F}(r|f) = \prod_{i=1}^N \boldsymbol{\sigma}((2r_i - 1)f_i)$ is not Gaussian, so the posterior is not Gaussian. But we can still calculate a Gaussian *approximation* $p_{F|R}(\cdot|r) \approx \mathcal{N}(\mu^{\text{p}}, \Sigma^{\text{p}})$. One based on Laplace approximation is given by (Lemma H.35)

$$\mu^{\text{p}} = f^\star \qquad\qquad \Leftrightarrow \qquad\qquad \mu^{\text{p}} = k(x)(r - \boldsymbol{\sigma}(f^\star))$$

$$\Sigma^{\text{p}} = (k(x)^{-1} + W^\star)^{-1} \qquad\qquad \Sigma^{\text{p}} = k(x) - k(x)\left((W^\star)^{-1} + k(x)\right)^{-1}k(x)$$

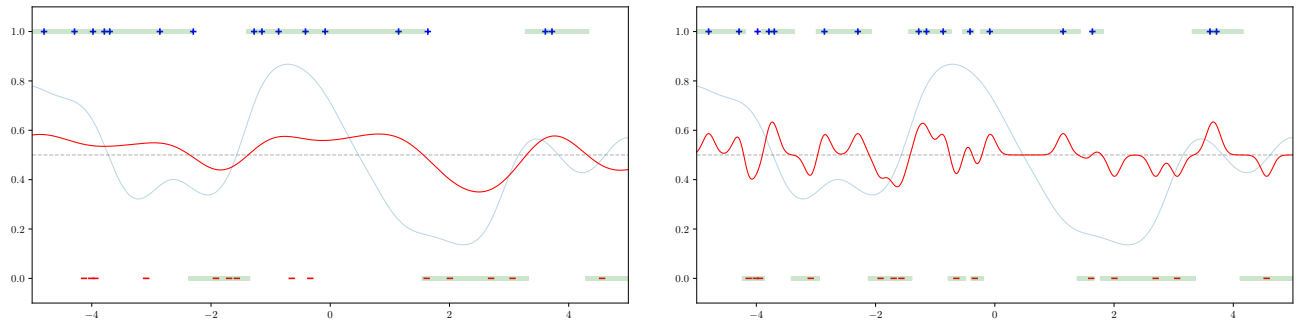


Figure 2: The ground-truth label probabilities are shown in blue lines. Training labels ($+ = 1$, $- = 0$) are generated according to these probabilities on random points. The expected probabilities (35) computed from the labeled points using (40) and (41) are shown in red lines, with the corresponding decision rules shown in green. The left plot uses the ground-truth bandwidth ($\sigma = 0.6$) and fails to fit the training data. The right plot uses a small bandwidth ($\sigma = 0.1$) to achieve 100% training accuracy.

where $f^\star \in \mathbb{R}^N$ is the unique maximizer of the strictly concave $\log p_{FR}(f, r)$ (hence can be recovered easily by numerical optimization, e.g., Newton's method) satisfying the stationary condition $f^\star = k(x)(r - \boldsymbol{\sigma}(f^\star))$, and $W^\star = \operatorname{diag}(\boldsymbol{\sigma}(f^\star) \odot (1 - \boldsymbol{\sigma}(f^\star)))$. The alternative form of $\Sigma^{\text{p}}$ is given by the Woodbury identity (65). Plugging these in (33) and (34), we have the Gaussian approximate predictive posterior $p_{F_{\text{test}}|R}(\cdot|r) \approx \mathcal{N}(\mu^{\text{pp}}, \Sigma^{\text{pp}})$ with

$$\mu^{\text{pp}} = k(x_{\text{test}}, x)(r - \boldsymbol{\sigma}(f^\star)) \tag{40}$$

$$\Sigma^{\text{pp}} = k(x_{\text{test}}) - k(x_{\text{test}}, x)(k(x) + (W^\star)^{-1})^{-1} k(x, x_{\text{test}}) \tag{41}$$

which can be used to calculate the expected probabilities (35) for classifying the test points. See Figure 2 for an illustration.

## 9.2 Sparse GPs

Standard GPs need to invert the $N \times N$ Gram matrix $k(x)$ which is $O(N^3)$. A mainstream approach to avoiding this computational difficulty is **sparse GPs**, which introduce additional unlabeled inputs $x_m \in \mathcal{X}^m$ where $m \ll N$ (e.g., by $k$-means on $x$). Now we have *three* latents, $(f_m, f, f_{\text{test}}) \in \mathbb{R}^{m+N+M}$ distributed as a GP on $(x_m, x, x_{\text{test}}) \in \mathcal{X}^{m+N+M}$. We can again easily compute the relations between any two latents because of joint normality, in particular

$$p_{F_{\text{test}}|F_m}(f_{\text{test}}|f_m) = \mathcal{N}\left(k(x_{\text{test}}, x_m)k(x_m)^{-1} f_m,\ k(x_{\text{test}}) - k(x_{\text{test}}, x_m)k(x_m)^{-1}k(x_m, x_{\text{test}})\right)(f_{\text{test}}) \tag{42}$$

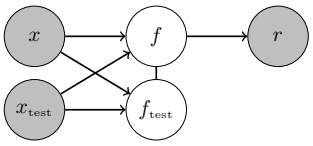
$$p_{F|F_m}(f|f_m) = \mathcal{N}\left(k(x, x_m)k(x_m)^{-1} f_m,\ k(x) - k(x, x_m)k(x_m)^{-1}k(x_m, x)\right)(f) \tag{43}$$

Note that they only require inverting the $m \times m$ Gram matrix $k(x_m)$. Letting $p_{R|F}(\cdot|f)$ denote the training likelihood, the predictive model under a sparse GP is

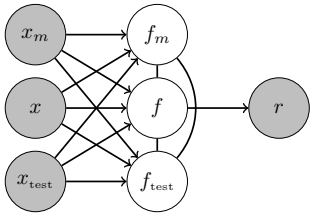
$$p_{F_m F F_{\text{test}} R}(f_m, f, f_{\text{test}}, r) = p_{F F_{\text{test}}}(f_m, f, f_{\text{test}}) \times p_{R|F}(r|f)$$

The difference between a standard GP and a sparse GP is illustrated below:

(Standard GP)




(Sparse GP)




Again, the goal is to estimate the predictive posterior $p_{F_{\text{test}}|R}(\cdot|r)$. In this case, it can be written as

$$p_{F_{\text{test}}|R}(f_{\text{test}}|r) = \int_{f_m \in \mathbb{R}^m} p_{F_m|R}(f_m|r) \times p_{F_{\text{test}}|F_m}(f_{\text{test}}|f_m)\, df \tag{44}$$

Since $p_{F_{\text{test}}|F_m}(\cdot|f)$ is Gaussian and known (42), if we also have a Gaussian form of the *sparse posterior* $p_{F_m|R}(\cdot|r) = \mathcal{N}(\nu^{\text{p}}, \Omega^{\text{p}})$ where $\nu^{\text{p}} \in \mathbb{R}^m$ and $\Omega^{\text{p}} \in \mathbb{R}^{m\times m}$ are functions of the training data, then (44) is again given by Bayes' rule (9) as $p_{F_{\text{test}}|R}(\cdot|r) = \mathcal{N}(\nu^{\text{pp}}, \Omega^{\text{pp}})$ where

$$\nu^{\text{pp}} = k(x_{\text{test}}, x_m)k(x_m)^{-1}\nu^{\text{p}} \tag{45}$$

$$\Omega^{\text{pp}} = k(x_{\text{test}}) - k(x_{\text{test}}, x_m)(k(x_m)^{-1} - k(x_m)^{-1}\Omega^{\text{p}}k(x_m)^{-1})k(x_m, x_{\text{test}}) \tag{46}$$

Critically, computing (45) and (46) can be done in $O(Mm^2 + m^3)$. We may then infer the test labels similarly as before. Since $\nu^{\text{p}}$ and $\Omega^{\text{p}}$ completely determine the solution, we only need to specify the Gaussian sparse posterior (or a Gaussian approximation of the sparse posterior) to make predictions.

### 9.2.1 Variational posterior approximation

We take a variational approach to recovering the sparse posterior $p_{F_m|R}(\cdot|r)$. The full posterior factorizes as $p_{F_mF|R}(f_m, f|r) = p_{F_m|R}(f_m|r) \times p_{F|F_m}(f|f_m)$ where the latter term is fully known (43). We can match the form of the approximate posterior by defining $q_{F_mF|R}(f_m, f|r) = q_{F_m|R}(f_m|r) \times p_{F|F_m}(f|f_m)$ so that we only need to optimize over $q_{F_m|R}$. The evidence lower bound (ELBO) on the MLL $\log p_R(r)$ becomes

$$\log p_R(r) \geq \mathop{\mathbf{E}}_{\substack{f_m\sim q_{F_m|R}(\cdot|r)\\ f|f_m\sim p_{F|F_m}(\cdot|f_m)}} \left[\log p_{R|F}(r|f)\right] - \text{KL}(q_{F_m|R}(\cdot|r), p_{F_m}) \tag{47}$$

By the usual property of the ELBO, a distribution $q_{F_m|R}(\cdot|r)$ that maximizes the bound satisfies $q_{F_m|R}(\cdot|r) = p_{F_m|R}(\cdot|r)$ and makes the inequality tight.

### 9.2.2 The regression sparse posterior

When the likelihood is Gaussian $p_{R|F}(\cdot|f) = \mathcal{N}(f, \Sigma)$, we have a closed-form solution of the true (Gaussian) sparse posterior and the corresponding MLL (Lemma H.36):

$$p_{F_m|R}(f_m|r) = \mathcal{N}\left(\Lambda(x_m)^{-1}k(x_m)^{-1}k(x_m, x)\Sigma^{-1}r, \Lambda(x_m)^{-1}\right)(f_m) \tag{48}$$

$$\log p_R(r) = \log \mathcal{N}(0_N, \Sigma + Q(x_m))(r) - \frac{1}{2}\text{tr}\left(\Sigma^{-1}\left(k(x) - Q(x_m)\right)\right) \tag{49}$$

where $\Lambda(x_m) = k(x_m)^{-1} + k(x_m)^{-1}k(x_m, x)\Sigma^{-1}k(x, x_m)k(x_m)^{-1}$ and $Q(x_m) = k(x, x_m)k(x_m)^{-1}k(x_m, x)$. Note that if $x_m = x$, then $\Lambda(x_m) = k(x)^{-1} + \Sigma^{-1}$ and $Q(x_m) = k(x)$ so that (48) and (49) reduce to

$$p_{F_m|R}(f_m|r) = \mathcal{N}\left((k(x)^{-1} + \Sigma^{-1})^{-1}\Sigma^{-1}r, (k(x)^{-1} + \Sigma^{-1})^{-1}\right)(f_m)$$

$$\log p_R(r) = \log \mathcal{N}(0_N, k(x) + \Sigma)(r)$$

which coincide with the posterior and the MLL in a standard GP for regression (Section 9.1.1). A notable feature of this result is that we can learn $x_m$ (aka. "inducing points") and any kernel hyperparameters by maximizing the MLL (49) (Titsias, 2009). The MLL can be computed in $O(Nm^2 + m^3)$ assuming a simple noise distribution (e.g., $\Sigma = \sigma^2 I_{N\times N}$); see this note for details on how to invert the covariance matrix efficiently. Once the desired variables in (49) are optimized, we compute the mean and covariance in (48) (again in $O(Nm^2 + m^3)$) for use in (45) and (46) to obtain $p_{F_{\text{test}}|R}(\cdot|r)$.

### 9.2.3 The classification sparse posterior

With the classification likelihood $p_{R|F}(r|f) = \prod_{i=1}^N \boldsymbol{\sigma}((2r_i - 1)f_i)$, the true sparse posterior $p_{F_m|R}(\cdot|r)$ is no longer Gaussian. But we can parameterize the approximate sparse posterior as Gaussian $q_{F_m|R}(\cdot|r) = \mathcal{N}(\nu^{\text{p}}, \Omega^{\text{p}})$ and directly optimize the ELBO over $\nu^{\text{p}} \in \mathbb{R}^m$ and $\Omega^{\text{p}} \in \mathbb{R}^{m\times m}$. (In practice, we reparameterize $\Omega^{\text{p}} = LL^\top$ where $L$ is a lower triangular matrix to enforce $\Omega^{\text{p}} \succ 0$ during optimization.) A few steps are needed to make this efficient (Hensman *et al.*, 2015). We marginalize out $f_m$ in $(f_m, f) \sim q_{F_m|R}(\cdot|r) \times p_{F|F_m}(\cdot|f_m)$ by Bayes' rule:

$$q_{F|R}(f|r) = \mathcal{N}\left(k(x, x_m)k(x_m)^{-1}\nu^{\text{p}}, k(x) + k(x, x_m)k(x_m)^{-1}\left(\Omega^{\text{p}} - k(x_m)\right)k(x_m)^{-1}k(x_m, x)\right)(f)$$

Note that $q_{F|R}(f|r) = \prod_{i=1}^N q_{F_i|R}(f_i|r)$ where

$$q_{F_i|R}(a|r) = \mathcal{N}\left(k(x_i, x_m)k(x_m)^{-1}\nu^{\text{p}}, k(x_i) + k(x_i, x_m)k(x_m)^{-1}\left(\Omega^{\text{p}} - k(x_m)\right)k(x_m)^{-1}k(x_m, x_i)\right)(a)$$

can be computed in $O(m^3)$ individually. The ELBO (47) becomes

$$\log p_R(r) \geq \sum_{i=1}^{N} \left( \int_{a \in \mathbb{R}} q_{F_i|R}(a|r) \times \log \boldsymbol{\sigma}((2r_i - 1)a)\, da \right) - \mathrm{KL}(\mathcal{N}(\nu^{\mathrm{p}}, \Omega^{\mathrm{p}}), \mathcal{N}(0_m, k(x_m))) \tag{50}$$

The Gaussian KL term is computed in $O(m^3)$ by (13). The one-dimensional integral of the log-likelihood can be estimated in $O(m^3)$ by Gaussian quadrature methods. The formulation (50) thus allows for scalable distributed optimization of $\nu^{\mathrm{p}}$ and $\Omega^{\mathrm{p}}$. Once they are optimized, we can compute the predictive posterior $p_{F_{\mathrm{test}}|R}$ and classify the test points by (35).

# 10 TODO: High-Dimensional Behavior

# References


Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Craig, A. T. (1943). Note on the independence of certain quadratic forms. *The Annals of Mathematical Statistics*, **14**(2), 195–197.

Driscoll, M. F. and Gundberg Jr, W. R. (1986). A history of the development of craig's theorem. *The American Statistician*, **40**(1), 65–70.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, **106**(496), 1602–1614.

Geary, R. (1936). The distribution of" student's" ratio for non-normal samples. *Supplement to the Journal of the Royal Statistical Society*, **3**(2), 178–184.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR.

Marsh, C. (2013). Introduction to continuous entropy. *Department of Computer Science, Princeton University*.

Petersen, K. B., Pedersen, M. S., *et al.* (2008). The matrix cookbook. *Technical University of Denmark*, **7**(15), 510.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Welling, M. (2013). Kernel ridge regression. *Max Welling's classnotes in machine learning*, pages 1–3.

# A Integration

## A.1 Single-Variable

An **antiderivative** of $f : \mathbb{R} \to \mathbb{R}$ is a function $F : \mathbb{R} \to \mathbb{R}$ such that $F' = f$. If $F$ is an antiderivative, then so is $F + C$ for any constant $C \in \mathbb{R}$. For instance, $(1/3)x^3 + 42$ is an antiderivative of $x^2$.

The (definite) **integral** of $f : \mathbb{R} \to \mathbb{R}$ over $[a, b]$ is a scalar $\int_a^b f(x)dx \in \mathbb{R}$ that represents the signed area of $f$ on $[a, b]$. The quantity $f(x)dx$ is interpreted as the product of the function value and an infinitesimally small interval. There are different ways to formalize the area. The most common definition is the Riemann integral which partitions $[a, b]$ into intervals $[i\delta, (i + 1)\delta]$ of width $\delta > 0$ and define

$$\int_a^b f(x)dx := \lim_{\delta \to 0} \sum_i f(x_i^\delta)\delta \tag{51}$$

where $x_i^\delta \in [i\delta, (i+1)\delta]$. The finite sum $\sum_i f(x_i^\delta)\delta$ for a given width $\delta$ is called a **Riemann sum**. Thus an integral is simply the limiting value of a Riemann sum (if it exists it is unique). A more general definition is a Lebesque integral which partitions the range of $f$.

The **fundamental theorem of calculus** (FTC) allows us to evaluate integrals by antiderivatives: if $F$ is any antiderivative of $f$, then

$$\int_a^b f(x)dx = F(x)\Big|_a^b := F(b) - F(a) \tag{52}$$

For instance, the signed area under $x^2$ over $[-1, 1]$ is $2/3$. Basic properties of integration include

$$\int_a^b \alpha f(x) + \beta g(x)dx = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx \qquad \text{(linearity)} \tag{53}$$

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du \qquad (u\text{-substitution}) \tag{54}$$

$$\int_a^b f(x)G(x)dx = F(x)G(x)\Big|_a^b - \int_a^b F(x)g(x)dx \qquad \text{(integration by parts)} \tag{55}$$

(Exercise: verify (54–55) using the chain rule and the product rule in differentiation.)

### A.1.1 Substitution in practice

While (54) is the standard form of $u$-substitution, we often use it mechanically as follows. We wish to integrate $f$ over the interval $a < b$. We view $f$ as a (hopefully simpler) function of $u = g(x)$ where $g : \mathbb{R} \to \mathbb{R}$ is invertible and differentiable with nonzero derivative over $(a, b)$. The infinitesimals are related as $du = g'(x)dx$ by the chain rule, or equivalently $dx = g'(g^{-1}(u))^{-1}du$. This yields a "plug-in" version of (54) where we substitute $g(x) = u$ and $dx = g'(g^{-1}(u))^{-1}du$,

$$\int_a^b f(g(x))dx = \int_{g(a)}^{g(b)} f(u)g'(g^{-1}(u))^{-1}du \tag{56}$$

For instance,

$$\begin{aligned}\int_0^{\sqrt{\frac{\pi}{2}}} 2x\cos\left(x^2\right)dx &= \int_0^{\frac{\pi}{2}} 2\sqrt{u}\cos(u)\left(\frac{1}{2\sqrt{u}}\right)du \\ &= \int_0^{\frac{\pi}{2}} \cos(u)\,du = \sin(u)\Big|_0^{\frac{\pi}{2}} = 1\end{aligned}$$

where $2x\cos\left(x^2\right) = 2\sqrt{u}\cos(u)$ with $u = g(x) = x^2$. Note that $g$ is invertible on $(0, \sqrt{\frac{\pi}{2}})$ so that $x = \sqrt{u}$; it is also differentiable with nonzero derivative $g'(x) = 2x$. Writing $dx = (2\sqrt{u})^{-1}du$, we cancel terms and are finally able to use FTC (52).

**Orientation of region.** Observe that

$$1 = \int_0^1 1dx = \int_0^{-1} (-1)du = \int_{-1}^0 (+1)du$$

The first equality is by FTC. The second equality is by (56) with $f(x) = 1$ and $u = g(x) = -x$. The final equality is again by FTC, simply acknowledging that $(-x)|_0^{-1} = x|_{-1}^0 = 1$. More generally, when $g'(x)^{-1} < 0$ (i.e., $u$ is moving in the opposite direction of $x$), we also change the "orientation of region" in integration (right-to-left instead of left-to-right). We can consider an alternative orientation-free formulation of $u$-substitution by always assuming integrating left-to-right. Let $R$ denotes a region $a < b$, then

$$\int_R f(g(x))dx = \int_{g(R)} f(u) \left|g'(g^{-1}(u))^{-1}\right| du \tag{57}$$

where $g(R)$ is the output region of $g$ when applied to $R$, integrated from a smaller value to a larger value. This formulation is useful because it generalizes to higher dimensions (59).

## A.2 Multi-Variable

The integral of $f : \mathbb{R}^d \to \mathbb{R}$ over a region $R \subseteq \mathbb{R}^d$ is a scalar $\int_R f(x)dx \in \mathbb{R}$ that represents the signed hypervolume of $f$ on $R$. Evaluation of such an integral is generally challenging because the region may take complicated forms (high-dimensional curves).

We can greatly simplify the problem by restricting the region to be a hypercube $R = [a, b]$ where $a, b \in \mathbb{R}^d$ specify a $d$-dimensional bounding box $[a_1, b_1] \times \cdots \times [a_d, b_d]$ (potentially all of $\mathbb{R}^d$). A central tool in this setting is **Fubini's thoerem**, which states that

$$\int_{[a,b]} f(x)dx = \int_{a_{\pi(d)}}^{b_{\pi(d)}} \left( \cdots \left( \int_{a_{\pi(1)}}^{b_{\pi(1)}} f(x_1 \dots x_d)dx_{\pi(1)} \right) \cdots \right) dx_{\pi(d)}$$

where $\pi$ is any permutation of $\{1 \dots d\}$. Thus we can evaluate a multi-variable integral by iteratively evaluating a single-variable integral in any order.

Many properties of integration carry over (like linearity), but some need to be generalized. One important generalization is **multi-variable $u$-substitution**. Let $R \subseteq \mathbb{R}^d$ and $g : R \to \mathbb{R}^d$ such that $J_g(x) \in \mathbb{R}^{d \times d}$ (Jacobian of $g$) is nonzero for all $x \in R$. Then

$$\int_R f(g(x)) \left|\det(J_g(x))\right| dx = \int_{g(R)} f(u)du \tag{58}$$

Similar to the single-variable case, we often use substitution mechanically as follows. We integrate $f$ over a region $R$ by viewing it as a simpler function of $u = g(x)$ where $g : R \to \mathbb{R}^d$ is assumed to be invertible (i.e., $\det(J_g(x)) \neq 0$). The infinitesimals are related as $du = \left|\det(J_g(x))\right| dx$ or equivalently $dx = \left|\det(J_g(x))\right|^{-1} du$. This gives

$$\int_R f(g(x))dx = \int_{g(R)} f(u) \left|\det(J_g(g^{-1}(u)))\right|^{-1} du \tag{59}$$

where we "plug in" $g(x) = u$ and $dx = \left|\det(J_g(g^{-1}(u)))\right|^{-1} du$. This strictly generalizes (57).

### A.2.1 Applications to probability

Let $X \in \mathbb{R}^d$ be a random vector with distribution $p_X$ supported on $S \subseteq \mathbb{R}^d$ (i.e., $p_X(x) \geq 0$ and $\int_S p_X(x)dx = 1$). The probability that $X$ lies in a region $R \subseteq S$ is

$$\Pr(X \in R) = \int_R p_X(x)dx$$

Let $t : S \to T$ be a smooth invertible function where $T \subseteq \mathbb{R}^d$. Define a new random vector $Y = t(X)$ supported on $T$. We claim that $Y$ has the distribution

$$p_Y(y) = p_X(t^{-1}(y)) \left|\det(J_{t^{-1}}(y))\right| \qquad \forall y \in T \tag{60}$$

Equivalently,

$$p_Y(t(x)) = p_X(x)\,|\det(J_{t^{-1}}(t(x)))| \qquad \forall x \in S \tag{61}$$

*Proof sketch.* For any $R \subseteq T$,

$$\Pr(Y \in R) = \Pr(X \in t^{-1}(R)) = \int_{t^{-1}(R)} p_X(x)dx = \int_R p_X(t^{-1}(y))\,|\det(J_{t^{-1}}(y))|\,dy$$

where the last equality applies (58) with $g = t^{-1}$. This implies (60).

# B Matrix Identities

**Inverse of a sum.** See Welling (2013) for (62) and (157) of Petersen *et al.* (2008) for (64).

$$(B^\top R^{-1} B + P^{-1})^{-1} B^\top R^{-1} = PB^\top (BPB^\top + R)^{-1} \qquad \text{(Welling)} \tag{62}$$

$$(R^{-1} + P^{-1})^{-1} R^{-1} = P(P + R)^{-1} \qquad \text{(Welling with } B = I\text{)} \tag{63}$$

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1} \qquad \text{(Woodbury)} \tag{64}$$

$$(A + B)^{-1} = A^{-1} - A^{-1}(B^{-1} + A^{-1})^{-1}A^{-1} \qquad \text{(Woodbury with } U = V = I\text{)} \tag{65}$$

**Block matrix inversion rule.** See the Wikipedia page.

$$\begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix}^{-1} = \begin{bmatrix} (M_1 - M_2 M_4^{-1} M_3)^{-1} & -(M_1 - M_2 M_4^{-1} M_3)^{-1} M_2 M_4^{-1} \\ -M_4^{-1} M_3 (M_1 - M_2 M_4^{-1} M_3)^{-1} & M_4^{-1} + M_4^{-1} M_3 (M_1 - M_2 M_4^{-1} M_3)^{-1} M_2 M_4^{-1} \end{bmatrix} \tag{66}$$

**Block matrix determinant rule.** See the Wikipedia page.

$$\det\left(\begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix}\right) = \det(M_1) \times \det(M_4 - M_3 M_1^{-1} M_2) \tag{67}$$

# C Continuous Entropy and KL Divergence

We generalize results in Marsh (2013) to multivariate. The continuous/differential entropy of $X \in \mathbb{R}^d$ with density $p_X$ supported on $S \subseteq \mathbb{R}^d$ is defined as[5]

$$H(X) := -\int_S p_X(x) \log p_X(x) dx \tag{68}$$

It is easily seen that entropy is additive for independent variables. That is, if $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are independent then the entropy of $Z = (X, Y) \in \mathbb{R}^{d+d'}$ is $H(Z) = H(X) + H(Y)$.

- The uniform distribution $u_{[a,b]}(x) := \frac{1}{b-a}$ over $[a, b] \subset \mathbb{R}$ has entropy

$$H(X) = \int_a^b \frac{1}{b-a} \log(b-a) dx = \log(b-a) \tag{69}$$

- The Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ over $\mathbb{R}^d$ has entropy (Corollary H.7)

$$H(X) = \frac{1}{2} \log\left((2\pi e)^d \det(\Sigma)\right)$$

- The exponential distribution $e_\lambda(x) := \lambda \exp(-\lambda x)$ over $[0, \infty)$ with parameter $\lambda > 0$ has entropy (Lemma H.5)

$$H(X) = 1 - \log \lambda \tag{70}$$

[5] We use the term "density" in this section to distingiush continuous vs discrete variables.

Unfortunately, continuous entropy suffers from various shortcomings (reviewed in Section C.1), most notably negativity (e.g., (69) is negative if $b-a<1$, (70) is negative if $\lambda > e$). On the other hand, let $q_X$ be another density of $X$ with support $S$. Define the continuous KL divergence (aka. relative entropy) between $p_X$ and $q_X$ as

$$\mathrm{KL}(p_X, q_X) := \int_S p_X(x) \log \frac{p_X(x)}{q_X(x)} dx \tag{71}$$

Continuous KL divergence is nonnegative:

$$\begin{aligned}
\mathrm{KL}(p_X, q_X) &= \mathop{\mathbf{E}}_{x\sim p_X}\left[\log \frac{p_X(x)}{q_X(x)}\right] \\
&= \mathop{\mathbf{E}}_{x\sim p_X}\left[-\log \frac{q_X(x)}{p_X(x)}\right] \\
&\geq -\log\left(\mathop{\mathbf{E}}_{x\sim p_X}\left[\frac{q_X(x)}{p_X(x)}\right]\right) && \text{(convexity of } -\log) \\
&= -\log\left(\int_S p_X(x)\frac{q_X(x)}{p_X(x)}dx\right) \\
&= -\log\left(\int_S q_X(x)dx\right) = 0
\end{aligned}$$

where $\mathrm{KL}(p_X, q_X) = 0$ iff $p_X = q_X$ almost everywhere. This has useful implications.

- The cross entropy between $p_X$ and $q_X$ upper bounds the entropy of $p_X$,

$$H(p_X, q_X) := H(p_X) + \mathrm{KL}(p_X, q_X) \geq H(p_X) \tag{72}$$

- Mutual information is nonnegative,

$$I(X, Y) := \mathrm{KL}(p_{XY}, p_X p_Y) \geq 0 \tag{73}$$

The cross entropy upper bound can be used to derive various maximum entropy densities.

**Theorem C.1.**

$$\mathcal{N}(\mu, \Sigma) \in \mathop{\arg\max}_{p_X:\, \mathbf{E}[X]=\mu,\, \mathrm{Var}(X)=\Sigma} H(p_X) \tag{74}$$

$$u_{[a,b]} \in \mathop{\arg\max}_{p_X:\, \mathrm{Support}(p_X)=[a,b]} H(p_X) \tag{75}$$

$$e_\lambda \in \mathop{\arg\max}_{p_X:\, \mathrm{Support}(p_X)=\mathbb{R}^d_{\geq 0},\, \mathbf{E}[X]=\lambda^{-1}} H(p_X) \tag{76}$$

where $u_{[a,b]}$ denotes the uniform distribution over $[a, b] \subset \mathbb{R}^d$ and $e_\lambda$ denotes the product exponential density over $\mathbb{R}^d_{\geq 0}$ with $\lambda > 0_d$

*Proof.* (74): Let $p_X$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \succ 0$. Then

$$\begin{aligned}
H(p_X, \mathcal{N}(\mu, \Sigma)) &= \int_{\mathbb{R}^d} p_X(x)\left(\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \frac{1}{2}\log((2\pi)^d \det(\Sigma))\right) \\
&= \frac{1}{2}\mathop{\mathbf{E}}_{x\sim p_X}\left[(x-\mu)^\top \Sigma^{-1}(x-\mu)\right] + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \\
&= \frac{d}{2} + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \\
&= \frac{1}{2}\log((2\pi e)^d \det(\Sigma)) = H(\mathcal{N}(\mu, \Sigma)) \geq H(p_X)
\end{aligned}$$

(75): Assume $d = 1$. Given any $p_X$ with support $[a, b]$ we have

$$H(p_X, u_X) = \int_a^b p_X(x)\log(b-a)dx = \log(b-a) = H(u_{[a,b]}) \geq H(p_X)$$

The statement holds for $d > 1$ since each dimension is independently optimized.

(76): Assume $d = 1$. Given any $p_X$ with support $[0, \infty)$ and mean $\lambda^{-1} > 0$ we have

$$H(p_X, e_\lambda) = \int_0^\infty p_X(x)(\lambda x - \log \lambda)dx = \lambda \mathop{\mathbf{E}}_{x\sim p_X}[x] - \log \lambda = 1 - \log \lambda = H(e_\lambda) \geq H(p_X, e_\lambda)$$

The statement holds for $d > 1$ since each dimension is independently optimized. □

## C.1 Shortcomings of Continuous Entropy

### C.1.1 Inconsistency with Shannon entropy

The Shannon entropy of discrete $X \in \{x_1 \dots x_n\}$ with distribution $p_X$ is

$$H(X) := -\sum_{i=1}^{n} p_X(x_i) \log p_X(x_i) \tag{77}$$

This definition was derived by Shannon as a solution that satisfies axioms of information (regarding monotonicity, non-negativity, zero information, and independence). (68) appears to be a natural continuous extension of (77) in the sense that both are $\mathbf{E}_{x\sim p_X}[-\log p_X(x)]$, but it fails to satisfy the axioms (e.g., it can be negative). One way to better understand why is to show that (68) is inconsistent with (77) in the limit. Assume $d = 1$ and let $p_X$ be a density supported on $[a, b]$. By definition

$$\int_a^b p_X(x)dx = \lim_{\delta\to 0} \sum_i p_X(x_i^\delta)\delta = 1 \tag{78}$$

where $\sum_i p_X(x_i^\delta)\delta$ is a finite Riemann sum of width $\delta > 0$. Thus we can cast the density $p_X$ as an increasingly fine-grained discrete distribution with probabilities $p_X(x_i^\delta)\delta$ as $\delta \to 0$. Note that each value of $\delta > 0$ yields a discrete distribution with a well-defined Shannon entropy. This Shannon entropy, in the limit, is

$$\begin{aligned}
\lim_{\delta\to 0}\left(-\sum_i (p_X(x_i^\delta)\delta)\log(p_X(x_i^\delta)\delta)\right) &= -\lim_{\delta\to 0}\sum_i (p_X(x_i^\delta)\log p_X(x_i^\delta))\delta - \lim_{\delta\to 0}\sum_i p_X(x_i^\delta)\delta\log\delta \\
&= -\int_a^b p_X(x)\log p_X(x)dx - \lim_{\delta\to 0}\sum_i p_X(x_i^\delta)\delta\log\delta \\
&= H(X) - \left(\lim_{\delta\to 0}\sum_i p_X(x_i^\delta)\delta\right)\left(\lim_{\delta\to 0}\log\delta\right) \qquad (79) \\
&= H(X) + \infty \qquad (80)
\end{aligned}$$

where (79) follows from the generalized product rule of limits using (78).[6] So the limiting Shannon entropy diverges from the continuous entropy by an infinite offset.

### C.1.2 Variability under change of coordinates

A good measure of information should not depend on the representation of samples from a distribution. For instance, let $p_X$ be a distribution over finitely many circles, each of which can be specified by its radius or area. Clearly, the Shannon entropy of the circle is the same regardless of the representation. Now let $p_X$ be a density over all circles. The continuous entropy of the circle under the radius representation is different from that under the area representation. A general statement that implies this result is given below.

**Lemma C.2.** Let $X \in \mathbb{R}^d$ with density $p_X$ supported on $S$. For any invertible mapping $t$ on $S$,

$$H(t(X)) = H(X) - \mathop{\mathbf{E}}_{x\sim p_X}[\log|\det(J_{t^{-1}}(t(x)))|]$$

[6] Assume $\lim_{x\to a} f(x) \neq 0$. If $g(x)$ does not oscillate around $a$,

$$\lim_{x\to a} f(x)g(x) = \lim_{x\to a} f(x) \lim_{y\to a} g(y)$$

If $g(x)$ oscillates around $a$, then so does $f(x)g(x)$.

*Proof.*

$$\begin{aligned} H(t(X)) &= -\int_S p_X(x) \log p_X(t(x)) dx \\ &= -\int_S p_X(x) \log p_X(x) dx - \int_S p_X(x) \log \left|\det(J_{t^{-1}}(t(x)))\right| dx && \text{(by (61))} \\ &= H(X) - \mathop{\mathbf{E}}_{x \sim p_X} \left[\log \left|\det(J_{t^{-1}}(t(x)))\right|\right] \end{aligned}$$

□

**Corollary C.3.** For any invertible $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$,

$$H(AX + b) = H(X) - \log \left|\det(A^{-1})\right| \tag{81}$$

**Corollary C.4.** For $\alpha > 0$,

$$H(\alpha X) = H(X) + d \log \alpha$$

*Proof.*

$$\begin{aligned} H(\alpha X) &= H(X) - \log \left|\det(\alpha^{-1} I_{d \times d})\right| && \text{(by (81))} \\ &= H(X) - \log \left|\alpha^{-d}\right| \\ &= H(X) - \log \alpha^{-d} && \text{(since } \alpha > 0\text{)} \\ &= H(X) + d \log \alpha \end{aligned}$$

□

Corollary C.4 states that we can vacuously increase the continuous entropy of $X \in \mathbb{R}^d$ to infinity by multiplying each value with a scalar $\alpha$ as we take $\alpha \to \infty$.

# D Moment-Generating Function

Let $X \in \mathbb{R}^d$ denote a random vector with distribution $p_X$. The **moment-generating function (MGF)** of $X$ is a real-valued positive mapping $M_X : \mathbb{R}^d \to (0, \infty)$ defined as

$$M_X(t) := \mathop{\mathbf{E}}_{x \sim p_X} \left[\exp\left(t^\top x\right)\right] \tag{82}$$

Not every distribution has a corresponding MGF (because (82) may diverge). But a classical result in probability theory is that an MGF uniquely determines a probability distribution. More formally, let $X, Y \in \mathbb{R}^d$ be random vectors with distributions $p_X, p_Y$ with well-defined MGFs $M_X, M_Y$. Then $p_X = p_Y$ iff $M_X = M_Y$. Thus an MGF is an alternative characterization of a random variable.

What makes $M_X$ special is obviously the exponential function. Since $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$,

$$M_X(t) = 1 + t^\top \underbrace{\mathbf{E}[X]}_{\text{1st moment}} + \frac{1}{2} t^\top \underbrace{\mathbf{E}\left[XX^\top\right]}_{\text{2nd moment}} t + \cdots$$

so that $\nabla^n M_X(0)$ is the $n$-th moment of $p_X$ (hence the name).

**Lemma D.1.** Let $X \sim \mathcal{N}(\mu, \Sigma)$. Then

$$M_X(t) = \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma t\right)$$

*Proof.* We use the same substitution in the proof of Lemma H.4. Let $\Sigma = U\Lambda U^\top$ denote an orthonormal eigendecomposition. Let $u = g(x)$ where $g(x) = \Lambda^{-1/2}U^\top(x-\mu)$, which implies $x = U\Lambda^{1/2}u + \mu$. Thus $|\det(J_g(x))| = \left|\det(\Lambda^{-1/2}U^\top)\right| = \det(\Lambda)^{-1/2}$, so we have the infinitesimal $dx = \sqrt{\det(\Lambda)}du$. Then

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\exp(t^\top x)dx \\
&= \int_{\mathbb{R}^d} \frac{\sqrt{\det(\Lambda)}}{(\sqrt{2\pi})^d\sqrt{\det(\Lambda)}} \exp\left(-\frac{1}{2}u^\top u\right)\exp(t^\top U\Lambda^{1/2}u + t^\top\mu)du \\
&= \exp(t^\top\mu)\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2}u^\top u + t^\top U\Lambda^{1/2}u\right)du \\
&= \exp(t^\top\mu)\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2}\left\|u - U\Lambda^{1/2}t\right\|^2 + \frac{1}{2}t^\top\Sigma t\right)du \\
&= \exp\left(t^\top\mu + \frac{1}{2}t^\top\Sigma t\right)\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2}\left\|u - U\Lambda^{1/2}t\right\|^2\right)du \\
&= \exp\left(t^\top\mu + \frac{1}{2}t^\top\Sigma t\right)
\end{aligned}$$

□

The first two moments of the Gaussian MGF are

$$h(t) := t^\top\mu + \frac{1}{2}t^\top\Sigma t$$

$$\nabla M_X(t) = \exp\left(h(t)\right)(\mu + \Sigma t) \qquad \Rightarrow \qquad \nabla M_X(0) = \mu = \mathbf{E}\left[X\right] \tag{83}$$

$$\nabla^2 M_X(t) = \exp\left(h(t)\right)\Sigma + \exp\left(h(t)\right)(\mu+\Sigma t)(\mu+\Sigma t)^\top \qquad \Rightarrow \qquad \nabla^2 M_X(0) = \Sigma + \mu\mu^\top = \mathbf{E}\left[XX^\top\right] \tag{84}$$

which imply that the mean and the covariance matrix of $X$ are $\mu$ and $\Sigma$.

An interesting consequence of Lemma D.1 is that a point-mass density can be viewed as a degenerate Gaussian distribution with zero variance. That is, if $X \in \mathbb{R}^d$ takes value $a \in \mathbb{R}^d$ with probability 1, then $M_X(t) = \exp(a^\top t)$, which is equal to the Gaussian MGF with $\Sigma = 0_{d\times d}$.

One application of MGF is showing that a linear transformation of a Gaussian random variable is also Gaussian. Note that the MGF of a linear transformation of $X$ is generally

$$M_{AX+b}(t) = \mathop{\mathbf{E}}_{x\sim p_X}\left[\exp\left(t^\top Ax\right)\exp\left(t^\top b\right)\right] = \exp\left(t^\top b\right)M_X(A^\top t) \tag{85}$$

**Lemma D.2.** Let $X \sim \mathcal{N}(\mu, \Sigma)$. Let $A \in \mathbb{R}^{d'\times d}$ and $b \in \mathbb{R}^{d'}$ where $d' \le d$ and $A$ has full rank. Then $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$.

*Proof.* For any $t \in \mathbb{R}^{d'}$,

$$\begin{aligned}
M_{AX+b}(t) &= \exp\left(t^\top b\right)M_X(A^\top t) && \text{(by (85))} \\
&= \exp\left(t^\top b\right)\exp\left(t^\top A\mu + \frac{1}{2}t^\top A\Sigma A^\top t\right) && \text{(by Lemma D.1)} \\
&= \exp\left(t^\top(A\mu + b) + \frac{1}{2}t^\top A\Sigma A^\top t\right)
\end{aligned}$$

The last term is the MGF of a random variable with distribution $\mathcal{N}(A\mu + b, A\Sigma A^\top)$ where $A\Sigma A^\top \succ 0$. The statement follows from the one-to-one correspondence between MGFs and distributions. □

## D.1 Cumulant-Generating Function

The log MGF $\psi_X(t) := \log \mathbf{E}[e^{t^\top X}]$ is called the **cumulant-generating function (CGF)** of $X$. We see that it is the (convex) log-partition function of $t$-tilted $X_t$ distributed as (Appendix E)

$$p_{X_t}(x) = \frac{e^{t^\top x}p_X(x)}{\mathbf{E}[e^{t^\top X}]}$$

We call $\nabla^{(n)}\psi_X(t)$ the $n$-th **cumulant** of $X$. From (93–94), we have

$$\nabla\psi_X(t) = \mathbf{E}[X_t] \tag{86}$$
$$\nabla^2\psi_X(t) = \mathrm{Cov}(X_t) \tag{87}$$

In particular,

$$\nabla\psi_X(0_d) = \mathbf{E}[X] \tag{88}$$
$$\nabla^2\psi_X(0_d) = \mathrm{Cov}(X) \tag{89}$$

This fact is used in Hoeffding's lemma which bounds the CGF of a bounded scalar random variable by using Taylor's approximation of the CGF around 0 and then bounding the mean/variance of that variable (Lemma H.27).

**Gaussian cumulants.** The CGF of $X \sim \mathcal{N}(\mu, \Sigma)$ is $\psi_X(t) = \mu^\top t + \frac{1}{2}t^\top \Sigma t$, so

$$\nabla\psi_X(t) = \mu + \Sigma t$$
$$\nabla^2\psi_X(t) = \Sigma$$

which is consistent with the fact that $X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma)$ (Lemma H.26). The corresponding Legendre transform $\psi_X^*(t) := \sup_{\lambda\in\mathbb{R}^d} \lambda^\top t - \psi_X(\lambda)$ of $\psi_X$ is (Lemma H.28)

$$\psi_X^*(t) = \frac{1}{2}(t-\mu)^\top \Sigma^{-1}(t-\mu) \tag{90}$$

# E Exponential Family

## E.1 Exponential Tilting

Given any "base" distribution $p$ over $\mathbb{R}^d$, we can generate a set of distributions $q_{p,\tau,\theta}$ by

$$q_{p,\tau,\theta}(x) := \frac{e^{\theta^\top \tau(x)} p(x)}{\mathbf{E}_{x'\sim p}[e^{\theta^\top \tau(x')}]} \tag{91}$$

for any $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ such that $\mathbf{E}_{x'\sim p}[e^{\theta^\top \tau(x')}]$ exists. Note that

- $q_{p,\tau,\theta}$ is nonnegative and sums/integrates to 1.
- $q_{p,\tau,\theta}$ has the same support as $p$.
- $q_{p,\tau,\theta}$ assigns a weight $e^{\theta^\top \tau(x)}$ on the probability of $x$, changing the tails of $p$.
- $q_{p,\tau,0_m} = p$.

This technique is called **exponential tilting** of $p$. We can rewrite (91) as

$$q_{p,\tau,\theta}(x) = p(x)\exp\left(\theta^\top \tau(x) - B_{p,\tau}(\theta)\right) \tag{92}$$

where the log-partition function $B_{p,\tau}(\theta) := \log \mathbf{E}_{x\sim p}[e^{\theta^\top \tau(x)}]$ normalizes $q_{p,\tau,\theta}$. We note several properties:

- $B_{p,\tau}(\theta)$ is convex (Lemma H.21).
- $\tau$ is a sufficient statistic for $\theta$ (Theorem H.20).
- Differentiating $B_{p,\tau}(\theta)$ generates the cumulants of $\tau(x)$ over $x \sim q_{p,\tau,\theta}$, for instance (Lemma H.22)

$$\nabla B_{p,\tau}(\theta) = \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)] \tag{93}$$
$$\nabla^2 B_{p,\tau}(\theta) = \mathop{\mathrm{Cov}}_{x\sim q_{p,\tau,\theta}}(\tau(x)) \tag{94}$$

  In particular, $\nabla B_{p,\tau}(0_m) = \mathbf{E}_{x\sim p}[\tau(x)]$ and $\nabla^2 B_{p,\tau}(0_m) = \mathrm{Cov}_{x\sim p}(\tau(x))$.
- Aside: (94) implies that $B_{p,\tau}(\theta)$ is convex since $\nabla^2 B_{p,\tau}(\theta) \succeq 0$.

Exponential tilting often preserves the distribution family. For instance, if $X \sim \mathcal{N}(\mu, \Sigma)$ and $X_t$ is the $t$-tilted $X$ with $t \in \mathbb{R}^d$ ($\tau(x) = x$), then $X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma)$ (Lemma H.26).

## E.2 Unnormalized Form

More generally, we may consider any nonnegative function $h : \mathbb{R}^d \to (0, \infty)$ ("base measure") and define

$$q_{h,\tau,\theta}(x) = \frac{\exp\left(\theta^\top \tau(x)\right) h(x)}{\int_{x\in\mathbb{R}^d} \exp\left(\theta^\top \tau(x)\right) h(x) dx} \tag{95}$$

for any $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ such that $\int_{x\in\mathbb{R}^d} \exp\left(\theta^\top \tau(x)\right) h(x) dx$ exists. We can rewrite (95) as

$$q_{h,\tau,\theta}(x) = h(x) \exp\left(\theta^\top \tau(x) - A_{h,\tau}(\theta)\right) \tag{96}$$

where $A_{h,\tau}(\theta) := \log\left(\int_{x\in\mathbb{R}^d} h(x) \exp\left(\theta^\top \tau(x)\right) dx\right)$ and $\tau$ is again a sufficient statistic for $\theta$. Clearly, exponential tilting is a special case where the base measure is normalized. However, (95) is strictly more general since it allows for $h$ such that $\int_x h(x)dx$ diverges. It is easy to check that the previous properties hold without a normalized base measure, specifically:

- Differentiating $A_{h,\tau}(\theta)$ generates the cumulants of $\tau(x)$ over $x \sim q_{h,\tau,\theta}$, in particular

$$\nabla A_{h,\tau}(\theta) = \mathop{\mathbf{E}}_{x\sim q_{h,\tau,\theta}} [\tau(x)] \tag{97}$$

$$\nabla^2 A_{h,\tau}(\theta) = \mathop{\mathrm{Cov}}_{x\sim q_{h,\tau,\theta}} (\tau(x)) \tag{98}$$

- (98) implies that $A_{h,\tau}(\theta)$ is convex.

A set of distributions that can be expressed in the form (96) is called an **exponential family**. $\theta \in \mathbb{R}^m$ is called its **natural parameter**. Note that there are many exponential families. For instance, the set of all normal distributions is one exponential family. The set of all categorical distributions is another exponential family.

### E.2.1 Discussions

**CGF.** The CGF $\psi_{\tau(X)}(t) = \log \mathbf{E}[e^{t^\top \tau(X)}]$ of $\tau(x)$ takes the form (Lemma H.24):

$$\psi_{\tau(X)}(t) = A_{h,\tau}(\theta + t) - A_{h,\tau}(\theta) \tag{99}$$

where we see $\nabla^{(n)} \psi_{\tau(X)}(0_m) = \nabla^{(n)} A_{h,\tau}(\theta)$; this is consistent with the fact that in an exponential family, the log-partition function generates cumulants.

**Conjugate prior.** In Bayesian probability theory, a prior over the parameter of a distribution is called a **conjugate prior** if the implied posterior over the parameter conditioning on a sample from the distribution is in the same distribution family that the prior is in. For an exponential family, we can define a prior

$$\pi_{h,\tau}(\theta; \alpha, \beta) = \frac{1}{Z_{h,\tau}(\alpha, \beta)} \exp\left(\theta^\top \alpha - \beta A_{h,\tau}(\theta)\right) \tag{100}$$

for any "pseudo-counts" $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}$ such that $Z_{h,\tau}(\alpha, \beta) = \int_{\theta\in\mathbb{R}^m} \exp\left(\theta^\top \alpha - \beta A_{h,\tau}(\theta)\right) d\theta$ exists. Then the posterior over $\theta$ given $x \sim q_{h,\tau,\theta}$ is given by (Lemma H.25)

$$\kappa_{h,\tau}(\theta | x; \alpha, \beta) = \pi_{h,\tau}(\theta; \tau(x) + \alpha, 1 + \beta) \tag{101}$$

thus (100) is a conjugate prior.

**Identifying an exponential family.** To check if a set of distributions $\left\{p(x; \bar{\theta})\right\}_{\bar{\theta}}$ is an exponential family, it is sufficient to propose any $h(x) \geq 0$, a transformation of $\bar{\theta}$ into natural parameter form $\theta = g(\bar{\theta}) \in \mathbb{R}^m$ and $x$ into sufficient statistic form $\tau(x) \in \mathbb{R}^m$, and *some* function $A_{h,\tau}(\theta)$, such that it can be written as (96):

$$p(x; \bar{\theta}) = q_{h,\tau,\theta}(x) = h(x) \exp\left(\theta^\top \tau(x) - A_{h,\tau}(\theta)\right)$$

In particular, we do not need to explicitly calculate $A_{h,\tau}(\theta) = \log\left(\int_{x\in\mathbb{R}^d} h(x) \exp\left(\theta^\top \tau(x)\right) dx\right)$ since the normalization of $p(x; \bar{\theta})$ enforces it (and guarantees its existence).

**Non-unique parameterization.** An exponential family has infinitely many equivalent parameterizations:

$$q_{h,\tau,\theta}(x) = q_{ah,u\odot\tau,\text{inv}(u)\odot\theta}(x) \qquad \forall a \in \mathbb{R}\setminus\{0\},\ u \in (\mathbb{R}\setminus\{0\})^m$$

where $\odot$ is the elementwise multiplication and $\text{inv}(u)$ is the elementwise inverse of vector $u$. It is often clear what a natural parameterization is (e.g., choose $u$ that makes $\tau(x)$ as simple as possible).

**Limitations.** A dizzying array of distributions are exponential families, including the normal (Lemma H.23), categorical, exponential, geometric, Bernoulli, Poisson, beta, and many others. But there are certain properties that an exponential family cannot capture. First, the form

$$h(x)\exp\left(\theta^\top\tau(x) - A_{h,\tau}(\theta)\right)$$

implies that the support of this distribution cannot depend on the parameter $\theta$. This rules out distributions like a uniform distribution on $[a,b] \subset \mathbb{R}$ whose support depends on the parameters $a, b$. Second, some distributions simply cannot be expressed using an inner product between the input and the parameter, for instance the Laplace distribution

$$\text{Laplace}(\mu,b)(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

Third, an exponential family necessarily has a well-defined MGF by (99), so it rules out distributions without an MGF such as the Cauchy distribution.

## E.3 Tweedie's Formula

**Lemma E.1.** Define the generative process over $t, x \in \mathbb{R}^d$:

$$t \sim p_T$$
$$x|t \sim p_{X|T}(\cdot|t) \qquad p_{X|T}(x|t) = \frac{e^{t^\top x}b(x)}{\mathbf{E}_{x\sim b}[e^{t^\top x}]}$$

where $b$ is some base distribution (i.e., we are exponential tilting it by $t$). Let

$$m(x) = \int_{t\in\mathbb{R}^d} p_T(t)p_{X|T}(x|t)dt$$

denote the marginal distribution over $x$. Define $l(x) := \log m(x)$ and $l_0(x) := \log b(x)$. We have

$$t|x \sim \mathbf{Unk}(\nabla l(x) - \nabla l_0(x),\ \nabla^2 l(x) - \nabla^2 l_0(x))$$

*Proof.* By Bayes' rule, the posterior over $t$ given $x$ is

$$p_{T|X}(t|x) = \frac{p_T(t)p_{X|T}(x|t)}{m(x)} = p_T(t)\mathbf{E}_{x\sim b}[e^{t^\top x}]\exp\left(t^\top x - \lambda(x)\right)$$

where $\lambda(x) = \log\frac{m(x)}{b(x)}$. This is an exponential family (96) with base measure $h(t) = p_T(t)\mathbf{E}_{x\sim b}[e^{t^\top x}]$, natural parameter $x \in \mathbb{R}^d$, sufficient statistic $t \in \mathbb{R}^d$, and the CGF $\lambda(x)$. By the usual property of the CFG, the mean $\mu$ and the covariance $\Sigma$ of $t \sim p_{T|X}(\cdot|x)$ is given by

$$\mu = \nabla\lambda(x) = \nabla l(x) - \nabla l_0(x)$$
$$\Sigma = \nabla^2\lambda(x) = \nabla^2 l(x) - \nabla^2 l_0(x)$$

□

**Lemma E.2** (Tweedie's formula)**.** Pick any $\Sigma \succ 0$. Define the generative process over $\mu, x \in \mathbb{R}^d$:

$$\mu \sim g$$
$$x|\mu \sim \mathcal{N}(\mu, \Sigma)$$

Let $m(x) = \int_{\mu \in \mathbb{R}^d} g(\mu)\mathcal{N}(\mu, \Sigma)(x)d\mu$ denote the marginal distribution over $x$. Define $l(x) := \log m(x)$. Then

$$\mu|x \sim \mathbf{Unk}(x + \Sigma\nabla l(x),\ \Sigma(I_{d\times d} + \nabla^2 l(x)\Sigma))$$

*Proof.* We can view $x|\mu \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N}(0_d + \Sigma t, \Sigma)$ as an exponential tilting of the base distribution $b(x) = \mathcal{N}(0_d, \Sigma)(x)$ by $t = \Sigma^{-1}\mu$ (Lemma H.26). Let $l_0(x) = \log b(x)$ and note that $\nabla l_0(x) = \nabla\left(-\frac{1}{2}x^\top\Sigma^{-1}x\right) = -\Sigma^{-1}x$. Lemma E.1 states that

$$\begin{aligned} t|x &\sim \mathbf{Unk}(\nabla l(x) - \nabla l_0(x),\ \nabla^2 l(x) - \nabla^2 l_0(x)) \\ &= \mathbf{Unk}(\Sigma^{-1}x + \nabla l(x),\ \Sigma^{-1} + \nabla^2 l(x)) \end{aligned}$$

Thus $\mu = \Sigma t$ conditioned on $x$ is distributed as

$$\mu|x \sim \mathbf{Unk}(x + \Sigma\nabla l(x),\ \Sigma + \Sigma\nabla^2 l(x)\Sigma)$$

□

# F Laplace Approximation

Let $p_Z$ denote a prior over $Z \in \mathbb{R}^d$ and $p_{X|Z}$ a likelihood over $X$ given $Z$. Conditioned on $X = x$, the **Laplace approximation** approximates the true posterior $p_{Z|X}(z|x) \propto p_{X|Z}(x|z) \times p_Z(z)$ by a Gaussian:

$$p_{Z|X}(z|x) \approx \mathcal{N}(z^\star,\ -H_x(z^\star)^{-1}) \tag{102}$$

where $l_x(z) = \log p_{Z|X}(z|x)$ is the log posterior, $z^\star = \arg\max_{z\in\mathbb{R}^d} l_x(z)$, and $H_x : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the Hessian of $l_x$ (which is assumed to exist). The approximation reflects the idea that the posterior is "pointy" around the mode and can be directly derived by the second-order Taylor approximation of $l_x$ around $z^\star$,

$$l_x(z^\star) \approx \frac{1}{2}(z - z^\star)^\top H_x(z^\star)(z - z^\star) + \text{constant}$$

which, when normalized, becomes the distribution $\mathcal{N}(z^\star, -H_x(z^\star)^{-1})$.

# G Linear Regression

Let $\boldsymbol{X} = (x_1 \dots x_N) \in \mathbb{R}^{N\times d}$ denote $N$ input vectors in matrix form, paired with continuous labels $y = (y_1 \dots y_N) \in \mathbb{R}^N$. In "generalized" least squares, we assume $y \sim \mathcal{N}(\boldsymbol{X}w^\star, \Sigma)$ for some unknown $w^\star \in \mathbb{R}^d$ and known positive-definite $\Sigma \in \mathbb{R}^{N\times N}$ (i.e., $y_i = w^\star \cdot x_i + \epsilon_i$ where $(\epsilon_1 \dots \epsilon_N) \sim \mathcal{N}(0_N, \Sigma)$). The hypothesis class is the family of conditional distributions $\mathcal{N}(\boldsymbol{X}w, \Sigma)$ over $\mathbb{R}^N$ indexed by $w \in \mathbb{R}^d$. The maximum-likelihood estimator (MLE) with an $l_2$ regularization coefficient $\lambda > 0$ is

$$\begin{aligned} \hat{w} &= \arg\max_{w\in\mathbb{R}^d} \log\left(\mathcal{N}(\boldsymbol{X}w, \Sigma)(y)\right) - \frac{\lambda}{2}||w||^2 \\ &= (\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{X} + \lambda I_{d\times d})^{-1}\boldsymbol{X}^\top\Sigma^{-1}y \\ &= \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\Sigma)^{-1}y \end{aligned} \tag{103}$$

where (103) uses the Welling identity (62). Given a test point $x_{\text{test}} \in \mathbb{R}^d$, the "true" label is produced by $y_{\text{test}} = w^\star \cdot x_{\text{test}} + \epsilon_{\text{test}}$ where $\epsilon_{\text{test}} \sim \mathcal{N}(0, \nu_{\text{test}})$. However, we predict $\hat{y} = \hat{w}^\top x_{\text{test}}$, plugging in $\hat{w}$ in place of $w^\star$ and assuming there is no test noise. Note that

$$\hat{y} = y^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\Sigma)^{-\top}\boldsymbol{X}x_{\text{test}}$$

is both (1) a linear combination of the training labels, and (2) a linear combination of the dot products between the test point and training inputs. The latter view admits the kernel trick: compute the Gram matrix $G \in \mathbb{R}^{N\times N}$ where $G_{i,j} = k(x_i, x_j)$ for a chosen kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (i.e., representing $\langle\phi(x_i), \phi(x_j)\rangle$ for some feature function $\phi$), compute $\alpha = (G + \lambda\Sigma)^{-1}y \in \mathbb{R}^N$, and predict $\hat{y} = \sum_{i=1}^N \alpha_i k(x_i, x_{\text{test}})$. Note that regularization is necessary to obtain the kernelized version (this is why it is called kernel *ridge* regression) and to invoke the representer theorem.

### G.1 Baysian Linear Regression

We assume $w \sim \mathcal{N}(0_d, \Omega)$ for some known positive-definite $\Omega \in \mathbb{R}^{d\times d}$. We now have a joint distribution $p(w, y|\boldsymbol{X}) = \mathcal{N}(0_d, \Omega)(w) \times \mathcal{N}(\boldsymbol{X}w, \Sigma)(y)$. By Gaussian Bayes' rule (Section 3.3), the associated marginal and posterior distributions are

$$\begin{aligned} p(y|\boldsymbol{X}) &= \mathcal{N}\left(0_N,\ \Sigma + \boldsymbol{X}\Omega\boldsymbol{X}^\top\right)(y) \\ \pi(w|\boldsymbol{X}, y) &= \mathcal{N}\left(\Omega\boldsymbol{X}^\top(\boldsymbol{X}\Omega\boldsymbol{X}^\top + \Sigma)^{-1}y,\ \Omega - \Omega\boldsymbol{X}^\top(\boldsymbol{X}\Omega\boldsymbol{X}^\top + \Sigma)^{-1}\boldsymbol{X}\Omega\right)(w) \end{aligned} \tag{104}$$

where (104) uses the Woodbury identity (64). We see that the ridge regressor (103) corresponds to the mode (also mean) of the posterior (104) using $\Omega = \lambda^{-1}I_{d\times d}$ (aka. the MAP estimate). Instead of using a single point, we can incorporate all of the posterior by considering the "predictive posterior":

$$\begin{aligned} p(\bar{y}|\boldsymbol{X}, y, x_{\text{test}}) &= \mathop{\mathbf{E}}_{w\sim\pi(\cdot|\boldsymbol{X},y)}\left[\mathcal{N}(w^\top x_{\text{test}}, \nu_{\text{test}})(\bar{y})\right] \\ &= \mathcal{N}\left(x_{\text{test}}^\top\Omega\boldsymbol{X}^\top(\boldsymbol{X}\Omega\boldsymbol{X}^\top + \Sigma)^{-1}y, \nu_{\text{test}} + x_{\text{test}}^\top\Omega x_{\text{test}} - x_{\text{test}}^\top\Omega\boldsymbol{X}^\top(\boldsymbol{X}\Omega\boldsymbol{X}^\top + \Sigma)^{-1}\boldsymbol{X}\Omega x_{\text{test}}\right)(\bar{y}) \end{aligned} \tag{105}$$

where the marginal (105) is again given by Bayes' rule. Assuming zero test noise $\nu_{\text{test}} = 0$ and defining the kernel function $k_\Omega(x, x') = x^\top\Omega x'$, we can express the predictive posterior as

$$\bar{y} \sim \mathcal{N}\left(k_\Omega(x_{\text{test}}, \boldsymbol{X})(k_\Omega(\boldsymbol{X}) + \Sigma)^{-1}y, k_\Omega(x_{\text{test}}) - k_\Omega(x_{\text{test}}, \boldsymbol{X})(k_\Omega(\boldsymbol{X}) + \Sigma)^{-1}k_\Omega(\boldsymbol{X}, x_{\text{test}})\right) \tag{106}$$

# H Lemmas

**Lemma H.1** (Polar coordinates)**.** For any integrable $f : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathbb{R}^2} f(x^2 + y^2)d(x, y) = 2\pi\int_0^\infty f(r^2)rdr$$

*Proof.* Let $R = [0, \infty) \times [0, 2\pi]$ and define $g : R \to \mathbb{R}^2$ by $g(r, \theta) = (r\cos\theta, r\sin\theta)$. Note that $r^2 = x^2 + y^2$ and $g(R) = \mathbb{R}^2$. The Jacobian of $g$ at $(r, \theta)$ is

$$J_g(r, \theta) = \begin{bmatrix} \frac{\partial r\cos\theta}{\partial r} & \frac{\partial r\cos\theta}{\partial\theta} \\ \frac{\partial r\sin\theta}{\partial r} & \frac{\partial r\sin\theta}{\partial\theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

Thus $|\det(J_g(r, \theta))| = \left|r(\cos^2\theta + \sin^2\theta)\right| = r$. Thus

$$\begin{aligned} \int_{\mathbb{R}^2} f(x^2 + y^2)d(x, y) &= \int_R f(g_1(r, \theta)^2 + g_2(r, \theta)^2)\left|J_g(r, \theta)\right| d(r, \theta) && \text{(by (58))} \\ &= \int_R f(r^2)rd(r, \theta) \\ &= \int_0^\infty\left(\int_0^{2\pi}\exp(-r^2)rd\theta\right)dr && \text{(Fubini)} \\ &= \int_0^\infty 2\pi\exp(-r^2)rdr && \text{(FTC)} \\ &= 2\pi\int_0^\infty\exp(-r^2)rdr && \text{(linearity)} \end{aligned}$$

□

**Lemma H.2** (Gaussian integral)**.**

$$\int_{-\infty}^\infty\exp\left(-x^2\right)dx = \sqrt{\pi} \tag{107}$$

*Proof.* A standard proof shows that $(\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx)^2 = \pi$ as follows:

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx\right)\left(\int_{-\infty}^{\infty} \exp\left(-y^2\right) dy\right) &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx\right) \exp\left(-y^2\right) dy && \text{(linearity)} \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} \exp\left(-x^2\right) \exp\left(-y^2\right) dx\right) dy && \text{(linearity)} \\
&= \int_{\mathbb{R}^2} \exp\left(-(x^2+y^2)\right) d(x,y) && \text{(Fubini)} \\
&= 2\pi \int_0^{\infty} \exp\left(-r^2\right) r dr && \text{(Lemma H.1)} \\
&= 2\pi \left(-\frac{1}{2}\exp(-r^2)\right)\Big|_0^{\infty} && \text{(FTC)} \\
&= 2\pi\left(0+\frac{1}{2}\right) = \pi
\end{aligned}$$

□

**Lemma H.3.** For any $\mu \in \mathbb{R}$ and $\sigma^2 > 0$,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = 1 \tag{108}$$

*Proof.* Let $u = \frac{x-\mu}{\sqrt{2}\sigma}$ which gives the infinitesimal $dx = \sqrt{2}\sigma du$. Then

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx &= \int_{-\infty}^{\infty} \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \exp\left(-u^2\right) du && \text{(by (56))} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-u^2\right) du \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-u^2\right) du && \text{(linearity)} \\
&= 1 && \text{(Lemma H.2)}
\end{aligned}$$

□

**Lemma H.4.**

$$\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) dx = 1$$

*Proof.* Let $\Sigma = U\Lambda U^\top$ denote an orthonormal eigendecomposition. Let $u = g(x)$ where $g(x) = \Lambda^{-1/2}U^\top(x-\mu)$. Thus $|\det(J_g(x))| = \left|\det(\Lambda^{-1/2}U^\top)\right| = \det(\Lambda)^{-1/2}$, so we have the infinitesimal $dx = \sqrt{\det(\Lambda)}du$. Then

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) dx &= \int_{\mathbb{R}^d} \frac{\sqrt{\det(\Lambda)}}{(\sqrt{2\pi})^d \sqrt{\det(\Lambda)}} \exp\left(-\frac{1}{2}u^\top u\right) du \\
&= \int_{\mathbb{R}^d} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du
\end{aligned}$$

By Fubini and linearity,

$$\begin{aligned}
\int_{\mathbb{R}^d} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du &= \int_{-\infty}^{\infty}\left(\cdots\left(\int_{-\infty}^{\infty} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du_1\right)\cdots\right) du_d \\
&= \prod_{i=1}^{d} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du_i \\
&= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx\right)^d = 1
\end{aligned}$$

where the last step applies Lemma H.3 with $\mu = 0$ and $\sigma^2 = 1$. □

**Lemma H.5.** For any $\lambda > 0$, the exponential distribution $e_\lambda(x) := \lambda \exp(-\lambda x)$ over $[0, \infty)$ has entropy

$$H(X) = 1 - \log \lambda$$

*Proof.*

$$\begin{aligned} H(X) &= -\int_0^\infty \lambda \exp(-\lambda x) \log(\lambda \exp(-\lambda x)) dx \\ &= -\log \lambda - \lambda \int_0^\infty \exp(-\lambda x)(-\lambda x) dx \end{aligned}$$

We evaluate the last integral as follows. Let $u = g(x) = -\lambda x$, then $g'(x) = -\lambda$ so that $\left|g'(g^{-1}(u))^{-1}\right| = 1/\lambda$. Reorienting the region between $g(0) = 0$ and $g(\infty) = -\infty$ and applying (57),

$$\begin{aligned} \lambda \int_0^\infty \exp(-\lambda x)(-\lambda x) dx &= \int_{-\infty}^0 \exp(u) u du \\ &= \exp(u) u|_{-\infty}^0 - \int_{-\infty}^0 \exp(u) du && \text{(integration by parts (55))} \\ &= (0 - 0) - \exp(u)|_{-\infty}^0 && (\lim_{u \to -\infty} \exp(u) u = 0) \\ &= -1 \end{aligned}$$

□

**Lemma H.6.** Define $\Delta := \mu' - \mu$. Then

$$H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \Delta^\top \Sigma^{-1} \Delta + \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \Sigma'\right) + \frac{1}{2} \log((2\pi)^d \det(\Sigma))$$

*Proof.*

$$\begin{aligned} H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) &:= \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} [-\log \mathcal{N}(\mu, \Sigma)(x)] \\ &= \frac{1}{2} \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[(x - \mu)^\top \Sigma^{-1} (x - \mu)\right] + \frac{1}{2} \log((2\pi)^d \det(\Sigma)) \end{aligned}$$

By the cyclic property and the linearity of trace,

$$\begin{aligned} \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[(x - \mu)^\top \Sigma^{-1} (X - \mu)\right] &= \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[\mathrm{tr}\left((x - \mu)^\top \Sigma^{-1} (x - \mu)\right)\right] \\ &= \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[\mathrm{tr}\left(\Sigma^{-1} (x - \mu)(x - \mu)^\top\right)\right] \\ &= \mathrm{tr}\left(\Sigma^{-1} \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[(x - \mu)(x - \mu)^\top\right]\right) \end{aligned}$$

Rewriting the expectation,

$$\begin{aligned} \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[(x - \mu)(x - \mu)^\top\right] &= \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[(x - \mu' + \Delta)(x - \mu' + \Delta)^\top\right] \\ &= \mathop{\mathbf{E}}_{x \sim \mathcal{N}(\mu', \Sigma')} \left[(x - \mu')(x - \mu')^\top + (x - \mu')\Delta^\top + \Delta(x - \mu')^\top + \Delta\Delta^\top\right] \\ &= \Sigma' + \Delta\Delta^\top \end{aligned}$$

Therefore we have

$$\begin{aligned} H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) &= \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \Sigma' + \Sigma^{-1} \Delta\Delta^\top\right) + \frac{1}{2} \log((2\pi)^d \det(\Sigma)) \\ &= \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \Sigma'\right) + \frac{1}{2} \Delta^\top \Sigma^{-1} \Delta + \frac{1}{2} \log((2\pi)^d \det(\Sigma)) \end{aligned}$$

□

**Corollary H.7** (Of Lemma H.6)**.**

$$H(\mathcal{N}(\mu,\Sigma)) = \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma)\right)$$

**Corollary H.8** (Of Lemma H.6 and Corollary H.7)**.** Define $\Delta := \mu' - \mu$. Then

$$\mathrm{KL}(\mathcal{N}(\mu',\Sigma'),\mathcal{N}(\mu,\Sigma)) = \frac{1}{2}\Delta^\top\Sigma^{-1}\Delta + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma' - I_{d\times d}\right) + \frac{1}{2}\log\frac{\det(\Sigma)}{\det(\Sigma')}$$

**Lemma H.9.** Let $A \in \mathbb{R}^{d\times d}$. The **main-diagonal block matrix** of $A$ at index $k \in \{1\dots d\}$ with size $n$ is a matrix $B(k,n) \in \mathbb{R}^{n\times n}$ with entries $B_{i,j}(k,n) = A_{k+i-1,k+j-1}$ for $i,j \in \{1\dots n\}$. If $A \succ 0$, then $B(k,n) \succ 0$ for all valid $k,n$.

*Proof.* Suppose $u^\top B(k,n)u \le 0$ for some nonzero $u \in \mathbb{R}^n$. Define $v \in \mathbb{R}^d$ where $v_{k+i-1} = u_i$ for $i = 1\dots n$ and other entries are zero. Then $v$ is nonzero and $v^\top Av = u^\top B(k,n)u \le 0$, contradicting the premise that $A \succ 0$. □

**Lemma H.10.** Let $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}_{\succ 0}$, $A \in \mathbb{R}^{d'\times d}$ and $\Omega \in \mathbb{R}^{d'\times d'}_{\succ 0}$. Then for all $y \in \mathbb{R}^{d'}$,

$$\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)}\left[\log\mathcal{N}(AX,\Omega)(y)\right] = \log\mathcal{N}(A\mu,\Omega)(y) - \frac{1}{2}\mathrm{tr}\left(\Omega^{-1}A\Sigma A^\top\right)$$

*Proof.* We have

$$\log\mathcal{N}(AX,\Omega)(y) = -\frac{1}{2}\log\left(2\pi\det(\Omega)^{d'}\right) - \frac{1}{2}(y-AX)^\top\Omega^{-1}(y-AX)$$

where

$$\begin{aligned}(y-AX)^\top\Omega^{-1}(y-AX) &= \mathrm{tr}\left((y-AX)^\top\Omega^{-1}(y-AX)\right)\\ &= \mathrm{tr}\left(\Omega^{-1}(y-AX)(y-AX)^\top\right)\\ &= \mathrm{tr}\left(\Omega^{-1}\left(yy^\top - yX^\top A^\top - AXy^\top + AXX^\top A^\top\right)\right)\end{aligned}$$

Since the trace is linear,

$$\begin{aligned}&\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)}\left[\mathrm{tr}\left(\Omega^{-1}\left(yy^\top - yX^\top A^\top - AXy^\top + AXX^\top A^\top\right)\right)\right]\\ &= \mathrm{tr}\left(\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)}\left[\Omega^{-1}\left(yy^\top - yX^\top A^\top - AXy^\top + AXX^\top A^\top\right)\right]\right)\\ &= \mathrm{tr}\left(\Omega^{-1}\left(yy^\top - y(A\mu)^\top - A\mu y^\top + A\left(\mu\mu^\top + \Sigma\right)A^\top\right)\right)\\ &= \mathrm{tr}\left(\Omega^{-1}\left(yy^\top - y(A\mu)^\top - A\mu y^\top + A\mu(A\mu)^\top + A\Sigma A^\top\right)\right)\\ &= \mathrm{tr}\left(\Omega^{-1}(y-A\mu)(y-A\mu)^\top\right) + \mathrm{tr}\left(\Omega^{-1}A\Sigma A^\top\right)\\ &= (y-A\mu)^\top\Omega^{-1}(y-A\mu) + \mathrm{tr}\left(\Omega^{-1}A\Sigma A^\top\right)\end{aligned}$$

Thus

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)}\left[\log\mathcal{N}(AX,\Omega)(y)\right] &= -\frac{1}{2}\log\left(2\pi\det(\Omega)^{d'}\right) - \frac{1}{2}(y-A\mu)^\top\Omega^{-1}(y-A\mu) - \frac{1}{2}\mathrm{tr}\left(\Omega^{-1}A\Sigma A^\top\right)\\ &= \log\mathcal{N}(A\mu,\Omega)(y) - \frac{1}{2}\mathrm{tr}\left(\Omega^{-1}A\Sigma A^\top\right)\end{aligned}$$

□

**Lemma H.11.** Let $X \sim \mathcal{N}(\mu,\Sigma)$. For any $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{m\times d}$,

$$A\Sigma B^\top = 0_{n\times m} \quad\Leftrightarrow\quad AX \in \mathbb{R}^n \text{ and } BX \in \mathbb{R}^m \text{ are independent}$$

*Proof.* If $A$ or $B$ is zero then the statement is trivially true (a constant is independent by definition). Otherwise, for all nonzero $(u, v) \in \mathbb{R}^{n+m}$, $(u, v)^\top (AX, BX) = (u^\top A + v^\top B)X$ is normal by the closure under linear transformation (2). Thus $(AX, BX)$ is normal by 4. Hence $AX$ and $BX$ are independent iff they are uncorrelated: $\mathbf{E}\left[A(X-\mu)(X-\mu)^\top B^\top\right] = A\Sigma B^\top = 0_{n\times m}$. □

**Lemma H.12.** Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be jointly normal with parameters $(\mu, \Sigma)$. Assume that $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible. Then for any $z = (x, y) \in \mathbb{R}^{d+d'}$,

$$\frac{1}{(\sqrt{2\pi})^{d+d'}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)\right) = \frac{1}{(\sqrt{2\pi})^{d}\sqrt{\det(\Sigma_X)}} \exp\left(-\frac{1}{2}(x-\mu_X)^\top \Sigma_X^{-1}(x-\mu_X)\right) \times \frac{1}{(\sqrt{2\pi})^{d'}\sqrt{\det(\Omega)}} \exp\left(-\frac{1}{2}(y-\phi(x))^\top \Omega^{-1}(y-\phi(x))\right) \tag{109}$$

where $\Omega \in \mathbb{R}^{d'\times d'}$ and $\phi(x) \in \mathbb{R}^{d'}$ are defined as

$$\Omega := \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} \tag{110}$$

$$\phi(x) := \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X) \qquad \forall x \in \mathbb{R}^d \tag{111}$$

*Proof.* By the block matrix inversion rule (66) and abbreviating $O = \Sigma_X^{-1}\Sigma_{XY}$,

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_X^{-1} + O\Omega^{-1}O^\top & -O\Omega^{-1} \\ -\Omega^{-1}O^\top & \Omega^{-1} \end{bmatrix}$$

Abbreviating $u = x - \mu_X$ and $v = y - \mu_Y$,

$$\begin{aligned}(z-\mu)^\top\Sigma^{-1}(z-\mu) &= u^\top\left(\Sigma_X^{-1} + O\Omega^{-1}O^\top\right)u - u^\top O\Omega^{-1}v - v^\top\Omega^{-1}O^\top u + v^\top\Omega^{-1}v \\ &= u^\top\Sigma_X^{-1}u + u^\top O\Omega^{-1}O^\top u - 2u^\top O\Omega^{-1}v + v^\top\Omega^{-1}v \\ &= u^\top\Sigma_X^{-1}u + (v - O^\top u)^\top\Omega^{-1}(v - O^\top u) \\ &= (x-\mu_X)^\top\Sigma_X^{-1}(x-\mu_X) + (y-\phi(x))^\top\Omega^{-1}(y-\phi(x))\end{aligned}$$

where we use the fact that $\Omega$ is symmetric. By the block matrix determinant rule (67), we have $\det(\Sigma) = \det(\Sigma_X)\det(\Omega)$. Applying these identities to the LHS of (109) yields the RHS. □

**Lemma H.13.** Let $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ where $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^{d'}$ and

$$\mu = \begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix} \qquad \Sigma = \begin{bmatrix}\Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y\end{bmatrix} \qquad \Lambda = \begin{bmatrix}\Lambda_X & \Lambda_{XY} \\ \Lambda_{YX} & \Lambda_Y\end{bmatrix} = \Sigma^{-1}$$

(i.e., $\Lambda$ is the precision matrix). Then

$$Y|X = x \sim \mathcal{N}\left(\mu_Y - \Lambda_Y^{-1}\Lambda_{YX}(x-\mu_X),\ \Lambda_Y^{-1}\right)$$

*Proof.* We can derive the log conditional probability of $Y = y$ given $X = x$ from the log joint probability $X = x, Y = x$ by treating all terms not involving $y$ as constants. Thus

$$\begin{aligned}\log\Pr(Y=y|X=x) &= -\frac{1}{2}((x,y)-\mu)^\top\Sigma^{-1}((x,y)-\mu) + C \\ &= -\frac{1}{2}((x,y)-\mu)^\top\Lambda((x,y)-\mu) + C \\ &= -\frac{1}{2}(y-\mu_Y)^\top\Lambda_Y(y-\mu_Y) - (x-\mu_X)^\top\Lambda_{XY}(y-\mu_Y) + C' \\ &= -\frac{1}{2}y^\top\Lambda_Y y + \left(\mu_Y^\top\Lambda_Y - (x-\mu_X)^\top\Lambda_{XY}\right)y + C''\end{aligned}$$

where the key step is directly expanding the precision matrix instead of inverse covariance. By matching the first- and second-order terms in Definition 5, we have $Y|X = x \sim \mathcal{N}(\nu(x), \Omega)$ where $\Omega = \Lambda_Y^{-1}$ and $\nu(x) = \Lambda_Y^{-1}(\Lambda_Y\mu_Y - \Lambda_{YX}(x-\mu_X)) = \mu_Y - \Lambda_{YX}(x-\mu_X)$. □

**Lemma H.14.** Let $X \sim \mathcal{N}(\mu, \Sigma_X)$ and $Y|X = x \sim \mathcal{N}(Ax + b, \Sigma_Y)$ for some $A \in \mathbb{R}^{d' \times d}$ and $b \in \mathbb{R}^{d'}$. Then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_X A^\top \\ A\Sigma_X & \Sigma_Y + A\Sigma_X A^\top \end{bmatrix} \right)$$

with the marginal and posterior distributions

$$\begin{aligned} Y &\sim \mathcal{N}(A\mu + b, \Sigma_Y + A\Sigma_X A^\top) \\ X|Y = y &\sim \mathcal{N}(\Lambda_X^{-1}(\Sigma_X^{-1}\mu + A^\top \Sigma_Y^{-1}(y - b)),\ \Lambda_X^{-1}) \qquad \Lambda_X^{-1} = (\Sigma_X^{-1} + A^\top \Sigma_Y^{-1} A)^{-1} \end{aligned}$$

*Proof.* With a bit of algebra, we can express the log probability of the variable $Z = (X, Y) \in \mathbb{R}^{d+d'}$ in the quadratic form:

$$\begin{aligned} -2 \log \Pr\left(Z = \begin{bmatrix} x \\ y \end{bmatrix}\right) &= -2 \log \Pr(X = x) - 2 \log \Pr(Y = y | X = x) \\ &= (x - \mu)^\top \Sigma_X^{-1}(x - \mu) + (y - Ax - b)^\top \Sigma_Y^{-1}(y - Ax - b) + C \\ &= x^\top (\Sigma_X^{-1} + A^\top \Sigma_Y^{-1} A)x + y^\top \Sigma_Y^{-1} y - y^\top \Sigma_Y^{-1} Ax - x^\top A^\top \Sigma_Y^{-1} y \\ &\quad + 2x^\top (A^\top \Sigma_Y^{-1} b - \Sigma_X^{-1}\mu) - 2y^\top \Sigma_Y^{-1} b + C' \\ &= \begin{bmatrix} x^\top & y^\top \end{bmatrix} \underbrace{\begin{bmatrix} \Sigma_X^{-1} + A^\top \Sigma_Y^{-1} A & -A^\top \Sigma_Y^{-1} \\ -\Sigma_Y^{-1} A & \Sigma_Y^{-1} \end{bmatrix}}_{\Lambda} \begin{bmatrix} x \\ y \end{bmatrix} + 2 \underbrace{\begin{bmatrix} A^\top \Sigma_Y^{-1} b - \Sigma_X^{-1}\mu \\ -\Sigma_Y^{-1} b \end{bmatrix}}_{-u} \begin{bmatrix} x \\ y \end{bmatrix} + C' \end{aligned}$$

which shows that the log probability of $Z = z$ is $-\frac{1}{2} z^\top \Lambda z + u^\top z + C''$ for some constant $C'' \in \mathbb{R}$. Thus $Z \sim \mathcal{N}(\nu, \Sigma)$ where $\Sigma = \Lambda^{-1}$ and $\nu = \Sigma u$ (Definition 5). By the block matrix inversion rule (66),

$$\Sigma = \Lambda^{-1} = \begin{bmatrix} \Sigma_X & -\Sigma_X(-A^\top \Sigma_Y^{-1})\Sigma_Y \\ -\Sigma_Y(-\Sigma_Y^{-1}A)\Sigma_X & \Sigma_Y + \Sigma_Y(-\Sigma_Y^{-1}A)\Sigma_X(-A^\top \Sigma_Y^{-1})\Sigma_Y \end{bmatrix} = \begin{bmatrix} \Sigma_X & \Sigma_X A^\top \\ A\Sigma_X & \Sigma_Y + A\Sigma_X A^\top \end{bmatrix}$$

Likewise, the mean is given by a lot of canceling terms:

$$\nu = \Sigma u = \begin{bmatrix} \Sigma_X & \Sigma_X A^\top \\ A\Sigma_X & \Sigma_Y + A\Sigma_X A^\top \end{bmatrix} \begin{bmatrix} \Sigma_X^{-1}\mu - A^\top \Sigma_Y^{-1} b \\ \Sigma_Y^{-1} b \end{bmatrix} = \begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}$$

This shows the statement about the marginal probability of $Y$. To obtain the statement about the posterior probability of $X$ given $Y = y$, we use the precision matrix form which states that (swapping $X$ and $Y$ in (8))

$$X|Y = x \sim \mathcal{N}\left(\mu_X - \Lambda_X^{-1}\Lambda_{XY}(y - \mu_Y),\ \Lambda_X^{-1}\right)$$

where $\mu_X = \mu$, $\mu_Y = A\mu + b$, $\Lambda_X = \Sigma_X^{-1} + A^\top \Sigma_Y^{-1} A$, and $\Lambda_{XY} = -A^\top \Sigma_Y^{-1}$. Noting $\Lambda_X = \Sigma_X^{-1} - \Lambda_{XY} A$, we can simplify the mean as

$$\mu - \Lambda_X^{-1}\Lambda_{XY}(y - A\mu - b) = \Lambda_X^{-1}(\Lambda_X \mu - \Lambda_{XY}(y - A\mu - b)) = \Lambda_X^{-1}(\Sigma_X^{-1}\mu + A^\top \Sigma_Y^{-1}(y - b))$$

□

**Lemma H.15.** Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be jointly normal with parameters $(\mu, \Sigma)$. Assume that $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible. Then for any $x \in \mathbb{R}^d$,

$$H(Y|X = x) = \frac{1}{2} \log\left((2\pi e)^{d'} \det(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right) \tag{112}$$

$$I(X, Y) = \frac{1}{2} \log\left(\frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)}\right) \tag{113}$$

*Proof.* By Lemma H.12, $Y|X=x$ is distributed as $\mathcal{N}(\phi(x),\Omega)$ for any $x\in\mathbb{R}^d$ where $\phi(x):=\mu_Y+\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)$ and $\Omega:=\Sigma_Y-\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$. Thus

$$H(Y|X=x)=\mathbf{E}\left[-\log\Pr(Y|X=x)\right]=\frac{1}{2}\mathbf{E}\left[(Y-\phi(x))^\top\Omega^{-1}(Y-\phi(x))\right]+\frac{1}{2}\log((2\pi)^{d'}\det(\Omega))$$

Using the cyclic property and linearity of trace, the first term is

$$\frac{1}{2}\mathbf{E}\left[(Y-\phi(x))^\top\Omega^{-1}(Y-\phi(x))\right]=\frac{1}{2}\mathrm{tr}\left(\Omega^{-1}\mathbf{E}\left[(Y-\phi(x))(Y-\phi(x))^\top\right]\right)=\frac{d'}{2}$$

This shows (112) (note the Euler constant $e$). To show (113), we have

$$\begin{aligned}I(X,Y)&=H(Y)-H(Y|X)\\&=\frac{1}{2}\log\left((2\pi e)^{d'}\det(\Sigma_Y)\right)-\frac{1}{2}\log\left((2\pi e)^{d'}\det(\Omega)\right)\\&=\frac{1}{2}\log\left(\frac{\det(\Sigma_Y)}{\det(\Omega)}\right)\\&=\frac{1}{2}\log\left(\frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)}\right)\end{aligned}$$

where for the last equality we use the fact that $\det(\Sigma)=\det(\Sigma_X\Omega)=\det(\Sigma_X)\det(\Omega)$. □

**Lemma H.16.** The following statements about $X\in\mathbb{R}^d$ are equivalent.

1. $X\sim\mathcal{N}(\mu,\Sigma)$, that is, $\Pr(X=x)=\frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)$.
2. $M_X(t)=\exp(t^\top\mu+\frac{1}{2}t^\top\Sigma t)$ for all $t\in\mathbb{R}^d$.
3. $X=\Sigma^{1/2}Z+\mu$ where $Z\sim\mathcal{N}(0_d,I_{d\times d})$.
4. $Y=a^\top X$ has the density $\mathcal{N}(a^\top\mu,a^\top\Sigma a)$ for all nonzero $a\in\mathbb{R}^d$.
5. $\log\Pr(X=x)=-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)+C=-\frac{1}{2}x^\top\Sigma^{-1}x+(\Sigma^{-1}\mu)^\top x+C'$ for some constants $C,C'\in\mathbb{R}$.

*Proof.* Lemma D.1 gives 1 ≡ 2. To show 2 ≡ 3 we note that by (85)

$$M_{\Sigma^{1/2}Z+\mu}(t)=\exp\left(t^\top\mu\right)M_Z(\Sigma^{1/2}t)=\exp\left(t^\top\mu+\frac{1}{2}t^\top\Sigma t\right)=M_X(t)$$

We have 1 ⇒ 4 since the density of $Y$ is $\mathcal{N}(a^\top\mu,a^\top\Sigma a)$ by Lemma D.2. To show 4 ⇒ 2, pick any nonzero $a\in\mathbb{R}^d$. For all $t\in\mathbb{R}$

$$M_X(ta)=M_{a^\top X}(t)=\exp\left(ta^\top\mu+\frac{1}{2}t^2a^\top\Sigma a\right)$$

where the first equality uses (85) and the second equality uses Lemma D.1. Setting $t=1$ gives $M_X(a)=\exp\left(a^\top\mu+\frac{1}{2}a^\top\Sigma a\right)$. Additionally, $M_X(0_d)=1=\exp\left(0_d^\top\mu+\frac{1}{2}0_d^\top\Sigma 0_d\right)$. Thus $M_X(t)=\exp(t^\top\mu+\frac{1}{2}t^\top\Sigma t)$ for all $t\in\mathbb{R}^d$. To show 1 ⇒ 5, the log probability of $X=x$ where $X\sim\mathcal{N}(\mu,\Sigma)$ is

$$\begin{aligned}\log\Pr(X=x)&=-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)+C_1\\&=-\frac{1}{2}x^\top\Sigma^{-1}x+\mu^\top\Sigma^{-1}x-\frac{1}{2}\mu^\top\Sigma^{-1}\mu+C_1&\text{(“expanding the square”)}\\&=-\frac{1}{2}x^\top\Sigma^{-1}x+(\Sigma^{-1}\mu)^\top x+C_2\end{aligned}$$

where $C_1,C_2\in\mathbb{R}$ are some constants independent of $x$. To show 5 ⇒ 1, note that

$$\begin{aligned}\log\Pr(X=x)&=-\frac{1}{2}x^\top\Sigma^{-1}x+(\Sigma^{-1}\mu)^\top x+C\\&=-\frac{1}{2}x^\top\Sigma^{-1}x+\mu^\top\Sigma^{-1}x-\frac{1}{2}\mu^\top\Sigma^{-1}\mu+\frac{1}{2}\mu^\top\Sigma^{-1}\mu+C\\&=-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)+C'&\text{(“completing the square”)}\end{aligned}$$

where $C'$ is a constant. This implies, for the constant $C'' = \exp(C')$,

$$\Pr(X = x) = C'' \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Since $\int_{x\in\mathbb{R}^d} \Pr(X = x)dx = 1$, we have

$$C'' = \frac{1}{\int_{x\in\mathbb{R}^d} \exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right) dx} = \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}}$$

where the second equality is by Lemma H.4. □

**Lemma H.17** (Popoviciu's inequality)**.** For any bounded scalar random variable $X \in [a, b]$,

$$\mathrm{Var}\,(X) \le \frac{(b-a)^2}{4}$$

with equality iff $\Pr(X = a) = \Pr(X = b) = \frac{1}{2}$.

*Proof.* For any constant $c \in \mathbb{R}$, $\mathbf{E}[(X-c)^2] = \mathbf{E}[(X - \mathbf{E}[X] + \mathbf{E}[X] - c)^2] \ge \mathrm{Var}\,(X)$. Choosing $c = \frac{b-a}{2}$ and using the fact that $\left|X - \frac{b-a}{2}\right| \le \frac{b-a}{2}$, we have $\mathrm{Var}\,(X) \le \mathbf{E}[(X - \frac{b-a}{2})^2] \le \frac{(b-a)^2}{4}$. □

**Lemma H.18** (Markov's inequality)**.** For any nonnegative scalar random variable $X \ge 0$, for any $\epsilon > 0$:

$$\Pr(X \ge \epsilon) \le \frac{\mathbf{E}[X]}{\epsilon}$$

*Proof.*

$$\begin{aligned}
\mathbf{E}[X] &= \int_0^\infty \Pr(X = x)x\, dx && \text{(proof similar if } X \text{ is discrete)}\\
&\ge \int_\epsilon^\infty \Pr(X = x)x\, dx\\
&\ge \int_\epsilon^\infty \Pr(X = x)\epsilon\, dx\\
&\ge \epsilon \Pr(X \ge \epsilon)
\end{aligned}$$

□

**Lemma H.19** (Chernoff's inequality)**.** For any scalar random variable $X \in \mathbb{R}$ and $\epsilon \ge \mathbf{E}[X]$,

$$\Pr(X \ge \epsilon) \le e^{-\psi_X^*(\epsilon)}$$

where $\psi_X^*(\epsilon) = \sup_{t\in\mathbb{R}} t\epsilon - \psi_X(t)$ is the Legendre transform of the CGF $\psi_X(t) = \log \mathbf{E}[e^{tX}]$.

*Proof.*

$$\begin{aligned}
\Pr(X \ge \epsilon) &\le \Pr(tX \ge t\epsilon) && \forall t \ge 0\\
&= \Pr(e^{tX} \ge e^{t\epsilon})\\
&= \frac{\mathbf{E}[e^{tX}]}{e^{t\epsilon}} && \text{(Markov's inequality, since } e^{tX} \ge 0 \text{ and } e^{t\epsilon} > 0)\\
&= e^{-(t\epsilon - \psi_X(t))}
\end{aligned}$$

In particular,

$$\begin{aligned}
\Pr(X \ge \epsilon) &\le \inf_{t\ge 0}\ e^{-(t\epsilon-\psi_X(t))}\\
&= e^{-\left(\sup_{t\ge 0}\ t\epsilon-\psi_X(t)\right)}\\
&= e^{-\left(\sup_{t\in\mathbb{R}}\ t\epsilon-\psi_X(t)\right)}\\
&= e^{-\psi_X^*(\epsilon)}
\end{aligned} \tag{114}$$

The step (114) uses the following lemma.

**Lemma.** Let $J(t) := t\epsilon - \psi_X(t)$ and $J^* = \sup_{t\in\mathbb{R}} J(t)$. Then $J^* \geq J(0)$.

**Proof.**

$$\begin{aligned} J(t) &= t\epsilon - \log \mathbf{E}[e^{tX}] \\ &\leq t\epsilon - t\mathbf{E}[X] && \text{(Jensen's inequality: } \log \mathbf{E}[X] \geq \mathbf{E}[\log X]\text{)} \\ &= t\underbrace{(\epsilon - \mathbf{E}[X])}_{\geq 0} \end{aligned}$$

Thus $J(t) \leq 0$ for all $t < 0$. The lemma follows from the fact that $J(0) = 0$.

□

**Theorem H.20** (Factorization Theorem)**.** Assume a joint distribution

$$p_{\Theta XT}(\theta, x, t) = p_\Theta(\theta) \times p_{X|\Theta}(x|\theta) \times [[\tau(x) = t]]$$

where $X \in \mathcal{X}$ is a sample from a distribution parameterized by $\Theta \in \mathcal{H}$, and $T = \tau(X) \in \mathcal{T}$ is the sample statistic for some function $\tau : \mathcal{X} \to \mathcal{T}$. The following statements about $\tau$ are equivalent: if any holds, we say $\tau$ is a **sufficient statistic** for $\Theta$.

- $X$ is conditionally independent of $\Theta$ given $T = t$:

$$p_{X|T}(x|t) = p_{X|T\Theta}(x|t, \theta) \tag{115}$$

- There exist $f_T : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{T} \times \mathcal{H} \to \mathbb{R}$ such that

$$p_{X|\Theta}(x|\theta) = f_T(x) \times g(\tau(x), \theta) \tag{116}$$

*Proof.* (116)⇒(115): For any $t, \theta$,

$$\begin{aligned} p_{T|\Theta}(t|\theta) &= \sum_{x\in\mathcal{X}:\ \tau(x)=t} p_{X|\Theta}(x|\theta) && \text{(proof similar if } X \text{ is continuous)} \\ &= \sum_{x\in\mathcal{X}:\ \tau(x)=t} f_T(x) \times g(\tau(x), \theta) && (116) \\ &= \left( \sum_{x\in\mathcal{X}:\ \tau(x)=t} f_T(x) \right) \times g(t, \theta) \end{aligned}$$

thus for any $x$ satisfying $\tau(x) = t$,

$$p_{X|T\Theta}(x|t, \theta) = \frac{p_{XT|\Theta}(x, t|\theta)}{p_{T|\Theta}(t|\theta)} = \frac{f_T(x) \times g(t, \theta)}{\left(\sum_{x\in\mathcal{X}:\ \tau(x)=t} f_T(x)\right) \times g(t, \theta)} = \frac{f_T(x)}{\sum_{x\in\mathcal{X}:\ \tau(x)=t} f_T(x)}$$

and $p_{X|T\Theta}(x|t, \theta) = 0$ for $x$ such that $\tau(x) \neq t$. This implies $p_{X|T}(x|t) = p_{X|T\Theta}(x|t, \theta)$ for all $\theta$.

(115)⇒(116): Define $f_T(x) = p_{X|T}(x|\tau(x))$ and $g(t, \theta) = p_{T|\Theta}(t|\theta)$. Then

$$\begin{aligned} p_{X|\Theta}(x|\theta) &= p_{XT|\Theta}(x, \tau(x)|\theta) \\ &= p_{X|T\Theta}(x|\tau(x), \theta) \times p_{T|\theta}(\tau(x)|\theta) \\ &= p_{X|T}(x|\tau(x)) \times p_{T|\Theta}(\tau(x)|\theta) && (115) \\ &= f_T(x) \times g(\tau(x), \theta) \end{aligned}$$

□

**Lemma H.21.** Let $X \in \mathcal{X}$ be a random variable and $\tau : \mathcal{X} \to \mathbb{R}^m$ be a function such that

$$B_{p,\tau}(\theta) := \log \mathbf{E}\left[e^{\theta^\top \tau(X)}\right]$$

exists for all $\theta \in \mathbb{R}^m$. Then $B_{p,\tau} : \mathbb{R}^m \to \mathbb{R}$ is convex.

*Proof.* We use Hölder's inequality which states that $\mathbf{E}[|XY|] \le \mathbf{E}[|X|^p]^{\frac{1}{p}}\mathbf{E}[|Y|^q]^{\frac{1}{q}}$ for any $p, q \ge 1$ satisfying $\frac{1}{p}+\frac{1}{q} = 1$. For any $\alpha \in [0,1]$ and $\theta, \omega \in \mathbb{R}^m$:

$$\begin{aligned}
\exp\left(B_{p,\tau}(\alpha\theta + (1-\alpha)\omega)\right) &= \mathbf{E}\left[e^{\alpha\theta^\top\tau(X)+(1-\alpha)\omega^\top\tau(X)}\right] \\
&= \mathbf{E}\left[\left|e^{\alpha\theta^\top\tau(X)}\right|\left|e^{(1-\alpha)\omega^\top\tau(X)}\right|\right] \\
&\le \mathbf{E}\left[\left|e^{\alpha\theta^\top\tau(X)}\right|^{\frac{1}{\alpha}}\right]^{\alpha}\mathbf{E}\left[\left|e^{(1-\alpha)\omega^\top\tau(X)}\right|^{\frac{1}{1-\alpha}}\right]^{(1-\alpha)} \qquad \left(p = \frac{1}{\alpha}, q = \frac{1}{1-\alpha}\right) \\
&= \mathbf{E}\left[e^{\theta^\top\tau(X)}\right]^{\alpha}\mathbf{E}\left[e^{\omega^\top\tau(X)}\right]^{(1-\alpha)} \\
&= \exp\left(B_{p,\tau}(\theta)\right)^{\alpha}\exp\left(B_{p,\tau}(\omega)\right)^{(1-\alpha)}
\end{aligned}$$

Taking the log on both sides yields $B_{p,\tau}(\alpha\theta + (1-\alpha)\omega) \le \alpha B_{p,\tau}(\theta) + (1-\alpha)B_{p,\tau}(\omega)$. □

**Lemma H.22.** Let $p$ be a distribution over $\mathbb{R}^d$ and define $q_{p,\tau,\theta}(x) := \frac{e^{\theta^\top\tau(x)}p(x)}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]}$ for function $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ where $\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]$ exists. Let $B_{p,\tau}(\theta) := \log \mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}]$. Then

$$\begin{aligned}
\nabla B_{p,\tau}(\theta) &= \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)] \\
\nabla^2 B_{p,\tau}(\theta) &= \mathop{\mathrm{Cov}}_{x\sim q_{p,\tau,\theta}}(\tau(x))
\end{aligned}$$

*Proof.*

$$\begin{aligned}
\nabla B_{p,\tau}(\theta) &= \frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]} \\
\nabla^2 B_{p,\tau}(\theta) &= \frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)\tau(x)^\top]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]} - \left(\frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]}\right)\left(\frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]}\right)^\top
\end{aligned}$$

Thus by the definition of $q_{p,\tau,\theta}$

$$\begin{aligned}
\nabla B_{p,\tau}(\theta) &= \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)] \\
\nabla^2 B_{p,\tau}(\theta) &= \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}\left[\tau(x)\tau(x)^\top\right] - \left(\mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)]\right)\left(\mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)]\right)^\top
\end{aligned}$$

□

**Lemma H.23.** $\mathcal{N}(\mu, \Sigma)$ is in the exponential family, with one parameterization given by

$$\begin{aligned}
h(x) &= \frac{1}{(\sqrt{2\pi})^d} && \text{(base measure)} \\
\theta &= \begin{bmatrix}\Sigma^{-1}\mu \\ -\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\end{bmatrix} \in \mathbb{R}^{d(d+1)} && \text{(natural parameter)} \\
\tau(x) &= \begin{bmatrix}x \\ \mathrm{vec}\left(xx^\top\right)\end{bmatrix} \in \mathbb{R}^{d(d+1)} && \text{(sufficient statistic)} \\
A_{h,\tau}(\theta) &= \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log\left(\det(\Sigma)\right)\right) && \text{(log-partition function)}
\end{aligned}$$

where $\mathrm{vec}(M) \in \mathbb{R}^{n^2}$ is the vector form of matrix $M \in \mathbb{R}^{n\times n}$ with $[\mathrm{vec}(M)]_{(i-1)n+j} = M_{i,j}$.

*Proof.*

$$\begin{aligned}\mathcal{N}(\mu,\Sigma)(x) &= \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\\ &= \frac{1}{(\sqrt{2\pi})^d}\exp\left(\mu^\top\Sigma^{-1}x-\frac{1}{2}x^\top\Sigma^{-1}x-\frac{1}{2}\mu^\top\Sigma^{-1}\mu-\frac{1}{2}\log\left(\det(\Sigma)\right)\right)\\ &= \frac{1}{(\sqrt{2\pi})^d}\exp\left(\begin{bmatrix}\Sigma^{-1}\mu\\ -\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\end{bmatrix}^\top\begin{bmatrix}x\\ \mathrm{vec}\left(xx^\top\right)\end{bmatrix}-\frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu+\log\left(\det(\Sigma)\right)\right)\right)\end{aligned}$$

where we use the fact that $u^\top Mv = \mathrm{vec}(M)^\top\mathrm{vec}(uv^\top)$. □

**Lemma H.24.** Let $q_{h,\tau,\theta}(x) = h(x)\exp\left(\theta^\top\tau(x) - A_{h,\tau}(\theta)\right)$ with $A_{h,\tau}(\theta) = \log\left(\int_{x\in\mathbb{R}^d} h(x)\exp\left(\theta^\top\tau(x)\right)dx\right)$ denote an exponential family. The log-MGF of the sufficient statistic $\tau(x)$ is given by

$$\psi_{\tau(X)}(t) = A_{h,\tau}(\theta+t) - A_{h,\tau}(\theta)$$

*Proof.*

$$\begin{aligned}M_{\tau(X)}(t) &= \mathop{\mathbf{E}}_{x\sim q_{h,\tau,\theta}}\left[\exp(t^\top\tau(x))\right]\\ &= \int_{x\in\mathbb{R}^d} h(x)\exp\left(\theta^\top\tau(x)-A_{h,\tau}(\theta)\right)\exp(t^\top\tau(x))dx\\ &= \exp\left(-A_{h,\tau}(\theta)\right)\int_{x\in\mathbb{R}^d} h(x)\exp\left((\theta+t)^\top\tau(x)\right)dx\\ &= \exp\left(A_{h,\tau}(\theta+t)-A_{h,\tau}(\theta)\right)\end{aligned}$$

□

**Lemma H.25.** Let $q_{h,\tau,\theta}(x) = h(x)\exp\left(\theta^\top\tau(x) - A_{h,\tau}(\theta)\right)$ with $A_{h,\tau}(\theta) = \log\left(\int_{x\in\mathbb{R}^d} h(x)\exp\left(\theta^\top\tau(x)\right)dx\right)$ denote an exponential family. Define a distribution over $\theta\in\mathbb{R}^m$ by

$$\pi_{h,\tau}(\theta;\alpha,\beta) := \frac{1}{Z_{h,\tau}(\alpha,\beta)}\exp\left(\theta^\top\alpha-\beta A_{h,\tau}(\theta)\right)$$

for $\alpha\in\mathbb{R}^m$ and $\beta\in\mathbb{R}$ such that $Z_{h,\tau}(\alpha,\beta) := \int_{\theta\in\mathbb{R}^m}\exp\left(\theta^\top\alpha-\beta A_{h,\tau}(\theta)\right)d\theta$ exists. Then the conditional distribution over $\theta$ given $x$ is

$$\kappa_{h,\tau}(\theta|x;\alpha,\beta) = \pi_{h,\tau}(\theta;\tau(x)+\alpha,1+\beta)$$

*Proof.* By Bayes' rule,

$$\begin{aligned}\kappa_{h,\tau}(\theta|x;\alpha,\beta) &\propto \pi_{h,\tau}(\theta;\alpha,\beta)\times q_{h,\tau,\theta}(x)\\ &= \frac{1}{Z_{h,\tau}(\alpha,\beta)}\exp\left(\theta^\top\alpha-\beta A_{h,\tau}(\theta)\right)\times h(x)\exp\left(\theta^\top\tau(x)-A_{h,\tau}(\theta)\right)\\ &\propto \exp\left(\theta^\top(\tau(x)+\alpha)-(1+\beta)A_{h,\tau}(\theta)\right)\end{aligned}$$

This implies $\kappa_{h,\tau}(\theta|x;\alpha,\beta) = \pi_{h,\tau}(\theta;\tau(x)+\alpha,1+\beta)$. □

**Lemma H.26.** Let $X_t$ denote the $t$-tilted $X\sim\mathcal{N}(\mu,\Sigma)$ using $\tau(x)=x$. Then

$$X_t\sim\mathcal{N}(\mu+\Sigma t,\Sigma)$$

*Proof.* We can directly verify this claim using the fact that the CGF of $X$ is $\mu^\top t + \frac{1}{2}t^\top \Sigma t$:

$$\begin{aligned}
\Pr(X_t = x) &= \frac{e^{t^\top x}}{\mathbf{E}_{x' \sim \mathcal{N}(\mu,\Sigma)}[e^{t^\top x'}]} \mathcal{N}(\mu,\Sigma)(x) \\
&= \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + t^\top x - \mu^\top t - \frac{1}{2}t^\top \Sigma t\right) \\
&= \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu-\Sigma t)^\top \Sigma^{-1}(x-\mu-\Sigma t)\right)
\end{aligned}$$

□

**Lemma H.27** (Hoeffding's lemma)**.** Let $X \in [a,b]$ be a bounded scalar random variable. Then

$$\psi_{X-\mathbf{E}[X]}(t) \le \frac{(b-a)^2 t^2}{8}$$

*Proof.* For any $t \in \mathbb{R}$, by Taylor's approximation of $\psi_X$ around 0, for some $\eta$ between 0 and $t$:

$$\psi_X(t) = \underbrace{\psi_X(0)}_{0} + \underbrace{\psi_X'(0)}_{\mathbf{E}[X]} t + \frac{1}{2}\underbrace{\psi_X''(\eta)}_{\mathrm{Var}(X_\eta)} t^2 \qquad \Leftrightarrow \qquad \psi_{X-\mathbf{E}[X]}(t) = \frac{\mathrm{Var}\left(X_\eta\right) t^2}{2}$$

where $X_\eta \in [a,b]$ is the $\eta$-tilted $X$ (87). By Popoviciu's inequality (Lemma H.17), $\mathrm{Var}\left(X_\eta\right) \le \frac{(b-a)^2}{4}$. □

**Lemma H.28.** Let $\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}^d} \lambda^\top t - \psi_X(\lambda)$ denote the Legendre transform of $\psi_X$. If $X \sim \mathcal{N}(\mu,\Sigma)$,

$$\psi_X^*(t) = \frac{1}{2}(t-\mu)^\top \Sigma^{-1}(t-\mu)$$

*Proof.* $J(\lambda) = \lambda^\top t - \psi_X(\lambda)$ is concave in $\lambda \in \mathbb{R}^d$ since $\psi_X$ is convex. The stationary condition is

$$\nabla J(\lambda) = t - \nabla \psi_X(\lambda) = t - \mu - \Sigma\lambda = 0_d$$

Thus $\lambda^* = \Sigma^{-1}(t-\mu)$ is the maximizer of $J$. Then

$$\begin{aligned}
\psi_X^*(t) &= (\lambda^*)^\top t - \psi_X(\lambda^*) \\
&= (\lambda^*)^\top t - (\lambda^*)^\top \mu - \frac{1}{2}(\lambda^*)^\top \Sigma \lambda^* \\
&= (t-\mu)^\top \Sigma^{-1} t - (t-\mu)^\top \Sigma^{-1}\mu - \frac{1}{2}(t-\mu)^\top \Sigma^{-1}(t-\mu) \\
&= \frac{1}{2}(t-\mu)^\top \Sigma^{-1}(t-\mu)
\end{aligned}$$

□

**Lemma H.29.** If $X \sim \mathcal{G}(\sigma^2)$, then $\mathrm{Var}\left(X\right) \le \sigma^2$.

*Proof.* By the Taylor series of $e^z = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots$,

$$\begin{aligned}
f(t) &:= \mathbf{E}[e^{tX}] = \mathbf{E}\left[1 + tX + \frac{t^2X^2}{2} + \frac{t^3X^3}{6} + \cdots\right] = 1 + \frac{t^2\mathbf{E}\left[X^2\right]}{2} + t^3 P_1(t) \\
g(t) &:= \mathbf{E}[e^{\frac{\sigma^2 t^2}{2}}] = 1 + \frac{\sigma^2 t^2}{2} + \frac{\sigma^4 t^4}{4} + \cdots = 1 + \frac{\sigma^2 t^2}{2} + t^3 P_2(t)
\end{aligned}$$

where $P_1, P_2$ are some polynomials. By premise, for all $t \in \mathbb{R}$

$$\begin{aligned}
f(t) \le g(t) \qquad &\Leftrightarrow \qquad \frac{t^2\mathbf{E}\left[X^2\right]}{2} + t^3 P_1(t) \le \frac{\sigma^2 t^2}{2} + t^3 P_2(t) \\
&\Leftrightarrow \qquad \mathbf{E}[X^2] - \sigma^2 \le tG(t)
\end{aligned}$$

where $G$ is again some polynomial. Thus

$$\mathbf{E}[X^2] - \sigma^2 \leq \lim_{t \to 0} tG(t) = 0 \qquad \Leftrightarrow \qquad \mathbf{E}[X^2] \leq \sigma^2$$

□

**Lemma H.30.** If $X, Z \in \mathbb{R}$ are random variables with the CGFs $\psi_X, \phi_Z : \mathbb{R} \to \mathbb{R}$,

$$\psi_X(t) \leq \phi_Z(t) \quad \forall t \in \mathbb{R} \qquad \Rightarrow \qquad \exp(-\psi_X^*(t)) \leq \exp(-\phi_Z^*(t)) \quad \forall t \in \mathbb{R}$$

where $\psi_X^*(t) = \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_X(t)$ is the Legendre transform of $\psi_X$ (similarly for $\psi_Z^*(t)$).

*Proof.*

$$\begin{aligned}
\psi_X(t) \leq \phi_Z(t) \quad &\Leftrightarrow \quad -\psi_X(t) \geq -\phi_Z(t) \\
&\Leftrightarrow \quad \lambda t - \psi_X(t) \geq \lambda t - \phi_Z(t) \qquad \forall \lambda \in \mathbb{R} \\
&\Rightarrow \quad \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_X(t) \geq \sup_{\lambda \in \mathbb{R}} \lambda t - \phi_Z(t) \\
&\Leftrightarrow \quad \psi_X^*(t) \geq \phi_Z^*(t) \\
&\Leftrightarrow \quad -\psi_X^*(t) \leq -\phi_Z^*(t) \\
&\Leftrightarrow \quad \exp(-\psi_X^*(t)) \leq \exp(-\phi_Z^*(t))
\end{aligned}$$

□

**Lemma H.31.** If $X_1 \dots X_N$ are independently sub-Gaussian with $X_i \sim \mathcal{G}(\sigma_i^2)$, then for all $\epsilon \geq 0$:

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{N^2\epsilon^2}{2\left(\sum_{i=1}^N \sigma_i^2\right)}\right)$$

*Proof.* Let $\widetilde{X} := \sum_{i=1}^N X_i$. Note that $\widetilde{X} \sim \mathcal{G}(\sum_{i=1}^N \sigma_i^2)$ (4) and $-\widetilde{X} \sim \mathcal{G}(\sum_{i=1}^N \sigma_i^2)$ (2). Thus

$$\begin{aligned}
\Pr\left(\left|\frac{1}{N}\widetilde{X}\right| \geq \epsilon\right) &= \Pr\left(\frac{1}{N}\widetilde{X} \leq -\epsilon \vee \frac{1}{N}\widetilde{X} \geq \epsilon\right) \\
&\leq \Pr\left(\frac{1}{N}\widetilde{X} \leq -\epsilon\right) + \Pr\left(\frac{1}{N}\widetilde{X} \geq \epsilon\right) && \text{(union bound)} \\
&= \Pr\left(-\widetilde{X} \geq N\epsilon\right) + \Pr\left(\widetilde{X} \geq N\epsilon\right) \\
&\leq 2\exp\left(-\frac{N^2\epsilon^2}{2\left(\sum_{i=1}^N \sigma_i^2\right)}\right) && (3)
\end{aligned}$$

□

**Lemma H.32.** The gradient $\nabla\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to \mathbb{R}^d$ and the Hessian $\nabla^2\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ of $\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to [0, 1]$ are

$$\begin{aligned}
\left(\nabla\mathcal{N}(\mu, \Sigma)\right)(x) &= -\mathcal{N}(\mu, \Sigma)(x) \times \Sigma^{-1}(x - \mu) \\
\left(\nabla^2\mathcal{N}(\mu, \Sigma)\right)(x) &= -\mathcal{N}(\mu, \Sigma)(x) \times \left(\Sigma^{-1} - \Sigma^{-1}(x - \mu)(x - \mu)^\top\Sigma^{-1}\right)
\end{aligned}$$

The Hessian is negative-definite at $x = \mu$, but possibly indefinite at other points.

*Proof.* Shorthanding $p(x) = \mathcal{N}(\mu, \Sigma)(x)$, $C = ((\sqrt{2\pi})^d\sqrt{\det(\Sigma)})^{-1}$, and $g(x) = -\frac{1}{2}(x - \mu)^\top\Sigma^{-1}(x - \mu)$ where $\nabla g(x) = -\Sigma^{-1}(x - \mu)$, we have

$$\nabla p(x) = C\nabla\exp\left(g(x)\right) = C\exp\left(g(x)\right)\nabla g(x) = p(x)\left(-\Sigma^{-1}(x - \mu)\right) = p(x)k(x)$$

with $k(x) = -\Sigma^{-1}(x-\mu)$. Denoting the Jacobian $J_k(x) = -\Sigma^{-1}$, we have

$$\nabla^2 p(x) = p(x)J_k(x) + k(x)\left(\nabla p(x)\right)^\top = -p(x)\Sigma^{-1} + p(x)k(x)k(x)^\top = -p(x)\left(\Sigma^{-1} - \Sigma^{-1}(x-\mu)(x-\mu)^\top\Sigma^{-1}\right)$$

Now we analyze the Hessian $H(x) = \nabla^2 p(x)$. To show that it is negative-definite at $x = \mu$, we simply note that

$$H(\mu) = -\underbrace{p(\mu)}_{>0}\underbrace{\Sigma^{-1}}_{\succ 0} \prec 0$$

(the inverse of a positive-definite matrix $\Sigma$ remains positive-definite). For the last statement, it is sufficient to give an example of an indefinite Hessian. Let $\mu = (0,0)$ and $\Sigma = I_{2\times 2}$ (i.e., standard Gaussian in $d = 2$ dimensions). Pick the point $x = (1,1)$, one standard deviation away from the mean in each dimension. For any vector $u \in \mathbb{R}^2$, we have

$$u^\top H(x)u = -p(x)\left(u^\top \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} u\right)$$

where the matrix is indefinite. For instance, $u = (1,-1)$ results in $-2p(x) < 0$ and $u = (1,1)$ results in $2p(x) > 0$. In the two-dimensional case, it is easy to visualize that this is a saddle point (convex along the direction of $(1,1)$, concave along the orthogonal direction). Interestingly, for $d = 1$, the points $x = \mu \pm \sigma$ result in $H(x) = 0$ ("inflection points"). □

**Lemma H.33.** $\frac{\partial}{\partial a}\Phi(\sqrt{\frac{\pi}{8}}a)\big|_{a=0} = \boldsymbol{\sigma}'(0) = \frac{1}{4}$

*Proof.* We have $\boldsymbol{\sigma}'(a) = \boldsymbol{\sigma}(a)(1-\boldsymbol{\sigma}(a))$ and $\boldsymbol{\sigma}(0) = \frac{1}{2}$, so $\boldsymbol{\sigma}'(0) = \frac{1}{4}$. On the other hand, we have

$$\frac{\partial \Phi(\lambda a)}{\partial a} = \lambda \mathcal{N}(0,1)(\lambda a) \qquad \Rightarrow \qquad \frac{\partial \Phi(\lambda a)}{\partial a}\bigg|_{a=0} = \lambda \mathcal{N}(0,1)(0) = \frac{\lambda}{\sqrt{2\pi}}$$

Matching the two values yield $\lambda = \sqrt{\frac{\pi}{8}}$. □

**Lemma H.34.** For any $\lambda, \beta \in \mathbb{R}$,

$$\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\Phi\left(\lambda X + \beta\right)\right] = \Phi\left(\frac{\lambda\mu + \beta}{\sqrt{1+\lambda^2\sigma^2}}\right)$$

*Proof.*

$$\begin{aligned}
\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\Phi\left(\lambda X + \beta\right)\right] &= \mathop{\mathbf{E}}_{Z\sim\mathcal{N}(0,1)}\left[\Phi\left(\lambda\mu + \lambda\sigma Z + \beta\right)\right] \\
&= \Pr_{Z,Z'\sim\mathcal{N}(0,1)}\left(Z' < \lambda\mu + \lambda\sigma Z + \beta\right) \\
&= \Pr_{Z,Z'\sim\mathcal{N}(0,1)}\left(Z' - \lambda\sigma Z < \lambda\mu + \beta\right) \\
&= \Pr_{Z''\sim\mathcal{N}(0,1+\lambda^2\sigma^2)}\left(Z'' < \lambda\mu + \beta\right) \\
&= \Pr_{Y\sim\mathcal{N}(0,1)}\left(Y < \frac{\lambda\mu+\beta}{\sqrt{1+\lambda^2\sigma^2}}\right) \\
&= \Phi\left(\frac{\lambda\mu+\beta}{\sqrt{1+\lambda^2\sigma^2}}\right)
\end{aligned}$$

Side note: proving a special case of this result (with $\beta = 0$) is an exercise (Exercise 4.26) in Bishop and Nasrabadi (2006), who give a very complicated problem-specific solution calculating integrals (which can be found online). In contrast, this proof is strikingly simple and well-known in the Stack Exchange community (e.g., here and here). This is a reminder that often the "right" solution is simple, and even the best people can miss it. □

**Lemma H.35.** In the GP classification model in Section 9.1.2, a Gaussian approximation of the posterior $\mathcal{N}(\mu^{\mathrm{p}}, \Sigma^{\mathrm{p}}) \approx p_{F|R}(\cdot|r)$ is given by

$$\begin{aligned}
\mu^{\mathrm{p}} &= f^\star \\
\Sigma^{\mathrm{p}} &= \left(k(x)^{-1} + W(f^\star)\right)^{-1}
\end{aligned}$$

where $f^\star \in \mathbb{R}^N$ is the unique maximizer of the strictly concave $\log p_{FR}(f,r)$ satisfying the stationary condition $f^\star = k(x)(r - \boldsymbol{\sigma}(f^\star))$, and $W(f) = \operatorname{diag}\left(\boldsymbol{\sigma}(f) \odot (1 - \boldsymbol{\sigma}(f))\right)$.

*Proof.* We apply the Laplace approximation (Appendix F) on

$$p_{FR}(f,r) = \mathcal{N}(0_N, k(x))(f) \times \left(\prod_{i=1}^N \boldsymbol{\sigma}((2r_i-1)f_i)\right)$$

to obtain $p_{F|R}(f|r) \approx \mathcal{N}(f^\star, -\nabla^2\tau_r(f^\star)^{-1})$ where $f^\star = \arg\max_f \log p_{F|R}(f|r) = \arg\max_f \log p_{FR}(f,r)$ and $\tau_r(f) = \log p_{F|R}(f|r)$ is the log posterior. With some calculation, we have[7]

$$\nabla\tau_r(f) = r - \boldsymbol{\sigma}(f) - k(x)^{-1}f$$
$$\nabla^2\tau_r(f) = -W(f) - k(x)^{-1} \prec 0$$

which shows $\tau_r$ is strictly concave and the unique optimum is obtained at a point $f^\star$ satisfying $f^\star = k(x)(r - \boldsymbol{\sigma}(f^\star))$. This gives the statement. □

**Lemma H.36.** With the Gaussian likelihood $p_{R|F}(r|f) = \mathcal{N}(f,\Sigma)(r)$ in a sparse GP, we have

$$p_{F_m|R}(f_m|r) = \mathcal{N}\left(\Lambda(x_m)^{-1}k(x_m)^{-1}k(x_m,x)\Sigma^{-1}r, \Lambda(x_m)^{-1}\right)(f_m)$$
$$\log p_R(r) = \log\mathcal{N}(0_N, \Sigma + Q(x_m))(r) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left(k(x) - Q(x_m)\right)\right)$$

where $\Lambda(x_m) = k(x_m)^{-1} + k(x_m)^{-1}k(x_m,x)\Sigma^{-1}k(x,x_m)k(x_m)^{-1}$ and $Q(x_m) = k(x,x_m)k(x_m)^{-1}k(x_m,x)$.

*Proof.* The ELBO (47) becomes

$$\log p_R(r) \geq \mathop{\mathbf{E}}_{\substack{f_m\sim q_{F_m|R}(\cdot|r)\\ f|f_m\sim\mathcal{N}(k(x,x_m)k(x_m)^{-1}f_m, k(x)-Q(x_m))}} [\log\mathcal{N}(f,\Sigma)(r)] - \mathrm{KL}(q_{F_m|R}(\cdot|r), p_{F_m})$$

Using (3), we can eliminate the latent $f$ in the ELBO:

$$\mathop{\mathbf{E}}_{f\sim\mathcal{N}(k(x,x_m)k(x_m)^{-1}f_m, k(x)-Q(x_m))}[\log\mathcal{N}(f,\Sigma)(r)] = \log\mathcal{N}(k(x,x_m)k(x_m)^{-1}f_m,\Sigma)(r) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left(k(x)-Q(x_m)\right)\right)$$

Combining the KL term, we can then write the ELBO as

$$\log p_R(r) \geq \mathop{\mathbf{E}}_{f_m\sim q_{F_m|R}(\cdot|r)}\left[\log\frac{G_{F_mR}(f_m,r)}{q_{F_m|R}(f_m|r)}\right] - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left(k(x)-Q(x_m)\right)\right) \tag{117}$$

where $G_{F_mR}(f_m,r) = p_{F_m}(f_m)\times\mathcal{N}(k(x,x_m)k(x_m)^{-1}f_m,\Sigma)(r)$ has a Gaussian prior and likelihood, thus its posterior is given by (10):

$$G_{F_m|R}(f_m|r) = \mathcal{N}\left(\Lambda^{-1}k(x_m)^{-1}k(x_m,x)\Sigma^{-1}r, \Lambda^{-1}\right)(f_m)$$

where $\Lambda = k(x_m)^{-1} + k(x_m)^{-1}k(x_m,x)\Sigma^{-1}k(x,x_m)k(x_m)^{-1}$.[8] Since $G_{F_mR}(f_m,r) \propto G_{F_m|R}(f_m|r)$ when $r$ is held fixed, for some $C_r$ (constant in $f_m$) we can write the ELBO as

$$\log p_R(r) \geq -\mathrm{KL}(q_{F_m|R}(\cdot|r), G_{F_m|R}(\cdot|r)) + C_r$$

[7] Here is the detail. Let $g(r) = \log p_{R|F}(r|f)$ and $h(f) = \log p_F(f)$. The first- and second-order partial derivatives of $g : \mathbb{R}^N \to \mathbb{R}$ are given by (letting $o = 2r - 1_N \in \{\pm1\}^N$ for convenience)

$$g(f) = \sum_{i=1}^N \log\boldsymbol{\sigma}(o_if_i) \qquad \frac{\partial g(f)}{\partial f_i} = o_i(1-\boldsymbol{\sigma}(o_if_i)) = r_i - \boldsymbol{\sigma}(f_i)$$
$$\frac{\partial^2 g(f)}{\partial f_i\partial f_j} = \begin{cases} -\boldsymbol{\sigma}(f_i)(1-\boldsymbol{\sigma}(f_i)) & \text{if } i=j \\ 0 & \text{otherwise}\end{cases}$$

Thus $\nabla g(f) = r - \boldsymbol{\sigma}(f)$ and $\nabla^2 g(f) = -W(f) \prec 0$. Turning to $h$, we have

$$h(f) = -\frac{1}{2}f^\top k(x)^{-1}f + \text{constant} \qquad \nabla h(f) = -k(x)^{-1}f$$
$$\nabla^2 h(f) = -k(x)^{-1} \prec 0$$

Since $\tau_r(f) = g(f) + h(f) + C_r$ for some $C_r$ constant in $f$, we have the statement.

[8] In the literature, an equivalent form is derived by convoluted algebraic manipulation. As a sanity check, we can convert to this form as follows. Let $S = k(x_m) + k(x_m,x)\Sigma^{-1}k(x,x_m)$ and note that $S = k(x_m)\Lambda k(x_m)$, thus $\Lambda^{-1} = k(x_m)S^{-1}k(x_m)$ and

$$\Lambda^{-1}k(x_m)^{-1}k(x_m,x)\Sigma^{-1}r = k(x_m)S^{-1}k(x_m,x)\Sigma^{-1}r \qquad \Lambda^{-1} = k(x_m)S^{-1}k(x_m)$$

The right-hand sides become the mean and covariance proposed in Titsias (2009) with $\Sigma = \sigma^2 I_{N\times N}$.

This shows that the optimal approximate posterior is $q^\star_{F_m|R} = G_{F_m|R}$. Since the search is not constrained, this must be the true posterior, thus $p_{F_m|R} = G_{F_m|R}$. This shows the first statement. For the second statement, while we should be able to plug $q^\star_{F_m|R}$ in the ELBO to derive the MLL, we get the result faster by noting that the first term in (117) is actually a "mini-ELBO" associated with $G_{F_m R}(f_m, r)$ and $q_{F_m|R}(f_m|r)$. Therefore, it attains its own maximum at $\log G_R(r)$ where $G_R(r) = \mathcal{N}(0_N, Q(x_m) + \Sigma)(r)$ is computed by (9). This shows

$$\log p_R(r) = \log \mathcal{N}(0_N, Q(x_m) + \Sigma)(r) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left(k(x) - Q(x_m)\right)\right)$$

□

# I Individually Normal But Not Jointly Normal

This is an example from Wikipedia. Let $X \sim \mathcal{N}(0,1)$ and, independently, $\epsilon \sim R$ where $R$ denotes the Rademacher distribution. Let $Y = \epsilon X$. By the symmetry of the distribution of $X$, we have $Y \sim \mathcal{N}(0,1)$. More formally,

$$\begin{aligned}
\Pr(Y \le x) &= \Pr(\epsilon = 1)\Pr(X \le x) + \Pr(\epsilon = -1)\Pr(X \ge -x) \\
&= \Pr(\epsilon = 1)\Pr(X \le x) + \Pr(\epsilon = -1)\Pr(-X \le x) \\
&= \frac{1}{2}\Pr(X \le x) + \frac{1}{2}\Pr(X \le x) \\
&= \Pr(X \le x)
\end{aligned}$$

Let $Z = X + Y$. Then $Z = 0$ with probability $\frac{1}{2}$ and $Z = 2X$ with probability $\frac{1}{2}$, so

$$\Pr(Z = z) = \frac{1}{2}\left([[z = 0]] + \mathcal{N}(0,1)\left(\frac{z}{2}\right)\right) \tag{118}$$

which is not a normal distribution. Then by definition 4, $(X, Y) \in \mathbb{R}^2$ is not normally distributed. Thus $X$ and $Y$ are not jointly normal, even though they are individually normal.

**Mutual information.** $X$ and $Y$ are uncorrelated. More formally,

$$\mathrm{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = \mathbf{E}\left[\epsilon X^2\right] = \mathbf{E}[\epsilon]\mathbf{E}\left[X^2\right] = 0$$

Thus $\mathrm{cor}(X, Y) = 0$. But $X$ and $Y$ are not independent. Specifically, $\Pr(Y = x|X = x) = \frac{1}{2}$ is not equal to $\Pr(Y = x) = \mathcal{N}(0,1)(x)$ for any $x \in \mathbb{R}$. This illustrates the limitation of linear correlation. On the other hand, the mutual information between $X$ and $Y$ is positive:

$$I(X, Y) = H(X) - H(X|Y) = H(X) - \log(2) = \log\sqrt{\frac{\pi e}{2}} \approx 0.73$$