The ELBO and the reparameterization trick

Ten problems on the evidence lower bound: its derivation from Jensen's inequality, the gap as a KL divergence to the posterior, the reconstruction-minus-KL form, the KL between diagonal Gaussians, the reparameterized gradients in μ, σ and log σ², the variance of the score-function estimator next to the reparameterized one, a linear Gaussian model solved in closed form with its tight bound, the one-sample gradient of a VAE step, and the Monte Carlo KL estimate and its variance, with worked solutions and the mistakes that reverse a KL, detach the sample or apply Jensen the wrong way.

BEFORE YOU START

A variational autoencoder is trained by maximizing a lower bound on the log-likelihood of the data, and the bound exists because of one inequality, Jensen's, applied to one expectation, over a distribution that the model chooses. Everything else is bookkeeping: the bound splits into a reconstruction term and a KL divergence, the KL between Gaussians has a closed form, and the gradient of the reconstruction term with respect to the distribution's own parameters is computed by writing the sample as a deterministic function of noise. These ten problems derive the evidence lower bound, identify the gap exactly, split it both ways, compute the Gaussian KL in the parameterization encoders use, derive the reparameterized gradient and compare its variance with the score-function alternative, solve a linear Gaussian model where every quantity is closed-form and the bound is tight, and work through the one-sample gradient of a VAE step and the one-sample KL estimate with its variance. The five mistakes at the end are the ones that train but train the wrong thing: the KL written in the other order, a sample treated as a constant, a Monte Carlo KL averaged under the prior, Jensen's inequality reversed, and a sample scaled by the variance instead of the standard deviation.

- A latent-variable model has a prior $p(z)$ over a latent z, a likelihood $p(x \mid z)$ for the observed x (in a VAE a decoder with parameters θ, written $p_\theta(x \mid z)$), the joint $p(x, z) = p(z)\,p(x \mid z)$, the evidence $p(x) = \int p(x, z)\,dz$ (a sum when z is discrete) and the posterior $p(z \mid x) = p(x, z)/p(x)$.
- A variational distribution $q(z)$ is any distribution over z that is positive wherever $p(x, z)$ is; in a VAE it is the encoder's output $q_\phi(z \mid x)$, and for a fixed x this page writes it $q(z)$. $\mathbb{E}_q$ is expectation under q. The evidence lower bound is $\mathcal{L}(q) = \mathbb{E}_q[\log p(x, z) - \log q(z)]$.
- Jensen's inequality for the concave logarithm: $\log \mathbb{E}[w] \geq \mathbb{E}[\log w]$ for a positive random variable w, with equality exactly when w is constant.

- From the entropy page: $\mathrm{KL}(q \,\|\, p) = \mathbb{E}_q[\log q(z) - \log p(z)] \geq 0$, with equality only when $q = p$; the entropy is $H(q) = -\mathbb{E}_q[\log q(z)]$; and for one-dimensional Gaussians, $\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma_1^2) \,\|\, \mathcal{N}(\mu_2, \sigma_2^2)\right) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$.
- A diagonal Gaussian $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ on $\mathbb{R}^d$ has independent coordinates $z_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Encoders output μ and s with $s_j = \log \sigma_j^2$, so $\sigma_j = e^{s_j/2}$.
- The reparameterization: $z = \mu + \sigma \odot \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$ has the distribution $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ (the multivariate-Gaussian page's Problem 8 with $L = \mathrm{diag}(\sigma)$). For $\varepsilon \sim \mathcal{N}(0, 1)$ the moments used below are $\mathbb{E}[\varepsilon] = \mathbb{E}[\varepsilon^3] = \mathbb{E}[\varepsilon^5] = 0$, $\mathbb{E}[\varepsilon^2] = 1$, $\mathbb{E}[\varepsilon^4] = 3$ and $\mathbb{E}[\varepsilon^6] = 15$.
- Variances and covariances follow the variance page: $\mathrm{Var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2$ and $\mathrm{Var}(u + v) = \mathrm{Var}(u) + \mathrm{Var}(v) + 2\,\mathrm{Cov}(u, v)$.

Builds on: Entropy, cross-entropy and KL divergence, The multivariate Gaussian: gradients and identities

Problems

1. Show that $\log p(x) \geq \mathcal{L}(q) = \mathbb{E}_q[\log p(x, z) - \log q(z)]$ for every admissible q, and say when the two are equal. difficulty ·

2. Show that $\log p(x) - \mathcal{L}(q) = \mathrm{KL}\left(q(z) \,\|\, p(z \mid x)\right)$. difficulty ·

3. Show that $\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid z)] - \mathrm{KL}\left(q(z) \,\|\, p(z)\right) = \mathbb{E}_q[\log p(x \mid z)] + \mathbb{E}_q[\log p(z)] + H(q)$. difficulty ··

4. Let $q = \mathcal{N}(\mu_1, \mathrm{diag}(\sigma_1^2))$ and $p = \mathcal{N}(\mu_2, \mathrm{diag}(\sigma_2^2))$ on $\mathbb{R}^d$. Show that difficulty ··

$$\mathrm{KL}(q \,\|\, p) = \sum_{j=1}^{d} \Big(\log \frac{\sigma_{2j}}{\sigma_{1j}} + \frac{\sigma_{1j}^2 + (\mu_{1j} - \mu_{2j})^2}{2\sigma_{2j}^2} - \frac{1}{2} \Big),$$

and that with $\mu_2 = 0$, $\sigma_2 = \mathbf{1}$ and $s_j = \log \sigma_{1j}^2$ it becomes $\frac{1}{2} \sum_j \left(\mu_{1j}^2 + e^{s_j} - s_j - 1\right)$.

5. Let $z = \mu + \sigma \odot \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, and $F(\mu, \sigma) = \mathbb{E}[f(z)]$ for a smooth $f : \mathbb{R}^d \to \mathbb{R}$. Show that $\nabla_\mu F = \mathbb{E}[\nabla f(z)]$, that $\partial F / \partial \sigma_j = \mathbb{E}[\partial_j f(z)\, \varepsilon_j]$, and that with $s_j = \log \sigma_j^2$, $\partial F / \partial s_j = \tfrac{\sigma_j}{2}\, \mathbb{E}[\partial_j f(z)\, \varepsilon_j]$. Verify all three on $f(z) = z^\top B z + c^\top z$ with B symmetric, where $F = \mu^\top B \mu + \sum_j B_{jj} \sigma_j^2 + c^\top \mu$. difficulty ··

6. Take $q = \mathcal{N}(\mu, \sigma^2)$ in one dimension and $f(z) = z^2$. Two unbiased estimators of $\partial\, \mathbb{E}_q[f(z)] / \partial \mu$ are the reparameterized $g_R = f'(\mu + \sigma\varepsilon) = 2(\mu + \sigma\varepsilon)$ and the score-function $g_S = f(z)\, \partial_\mu \log q(z) = z^2 (z - \mu)/\sigma^2$ with $z = \mu + \sigma\varepsilon$. Show both have mean 2μ and compute their variances. difficulty ···

7. A linear Gaussian model: $p(z) = \mathcal{N}(0, 1)$, $p(x \mid z) = \mathcal{N}(wz + b, \gamma^2)$ with w, b, γ fixed, and $q(z) = \mathcal{N}(\mu, \sigma^2)$. Show that difficulty ··

$$\begin{aligned} \mathcal{L}(\mu, \sigma^2) = &-\tfrac{1}{2} \log(2\pi\gamma^2) \\ &- \frac{(x - w\mu - b)^2 + w^2 \sigma^2}{2\gamma^2} \\ &- \tfrac{1}{2}\big(\mu^2 + \sigma^2 - \log \sigma^2 - 1\big). \end{aligned}$$

8. Maximize Problem 7's $\mathcal{L}$ over μ and σ^2. Show that $\mu^* = \dfrac{w(x - b)}{\gamma^2 + w^2}$ and $\sigma^{*2} = \dfrac{\gamma^2}{\gamma^2 + w^2}$, that $q^* = \mathcal{N}(\mu^*, \sigma^{*2})$ is the exact posterior $p(z \mid x)$, and that the bound is tight: $\mathcal{L}^* = \log p(x)$ with $p(x) = \mathcal{N}(x; b, w^2 + \gamma^2)$. difficulty ···

9. Let the decoder give $\log p_\theta(x \mid z) = g(z)$ for a differentiable g, with $q = \mathcal{N}(\mu, \sigma^2)$ and prior $\mathcal{N}(0, 1)$. The one-sample estimate of the ELBO is $\hat{\mathcal{L}}(\mu, \sigma) = g(\mu + \sigma\varepsilon) - \tfrac{1}{2}\big(\mu^2 + \sigma^2 - \log \sigma^2 - 1\big)$ with ε drawn once. Compute $\partial \hat{\mathcal{L}} / \partial \mu$, $\partial \hat{\mathcal{L}} / \partial \sigma$ and $\partial \hat{\mathcal{L}} / \partial s$ with $s = \log \sigma^2$, and show that each has expectation equal to the corresponding gradient of $\mathcal{L}$. difficulty ···

10. Instead of Problem 4's closed form, the KL can be estimated from one sample: $\hat{K} = \log q(z) - \log p(z)$ with $z = \mu + \sigma\varepsilon$, $q = \mathcal{N}(\mu, \sigma^2)$ and $p = \mathcal{N}(0, 1)$. Show that $\hat{K} = -\log \sigma + \tfrac{1}{2}\mu^2 + \mu\sigma\varepsilon + \tfrac{1}{2}(\sigma^2 - 1)\varepsilon^2$, that $\mathbb{E}[\hat{K}] = \mathrm{KL}(q \,\|\, p)$, and that $\mathrm{Var}(\hat{K}) = \mu^2\sigma^2 + \tfrac{1}{2}(\sigma^2 - 1)^2$. difficulty ···

Answers

1. $\log p(x) \geq \mathcal{L}(q) = \mathbb{E}_q[\log p(x, z) - \log q(z)]$, with equality exactly when $q(z) = p(z \mid x)$
2. $\log p(x) = \mathcal{L}(q) + \mathrm{KL}\left(q(z) \,\|\, p(z \mid x)\right)$, so $\log p(x) - \mathcal{L}(q) = \mathrm{KL}\left(q \,\|\, p(z \mid x)\right) \geq 0$
3. $\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid z)] - \mathrm{KL}\left(q(z) \,\|\, p(z)\right) = \mathbb{E}_q[\log p(x \mid z)] + \mathbb{E}_q[\log p(z)] + H(q)$
4. $\mathrm{KL}(q \,\|\, p) = \sum_j \left(\log \dfrac{\sigma_{2j}}{\sigma_{1j}} + \dfrac{\sigma_{1j}^2 + (\mu_{1j} - \mu_{2j})^2}{2\sigma_{2j}^2} - \dfrac{1}{2} \right)$; against $\mathcal{N}(0, I)$ with $s_j = \log \sigma_{1j}^2$ it is $\frac{1}{2} \sum_j \left(\mu_{1j}^2 + e^{s_j} - s_j - 1 \right)$
5. $\nabla_\mu F = \mathbb{E}[\nabla f(z)]$, $\dfrac{\partial F}{\partial \sigma_j} = \mathbb{E}[\partial_j f(z)\, \varepsilon_j]$, $\dfrac{\partial F}{\partial s_j} = \dfrac{\sigma_j}{2}\, \mathbb{E}[\partial_j f(z)\, \varepsilon_j]$; for the quadratic both routes give $2B\mu + c$, $2B_{jj}\sigma_j$ and $B_{jj}\sigma_j^2$
6. Both estimators have mean 2μ; $\mathrm{Var}(g_R) = 4\sigma^2$ while $\mathrm{Var}(g_S) = \dfrac{\mu^4}{\sigma^2} + 14\mu^2 + 15\sigma^2$
7. $\mathcal{L}(\mu, \sigma^2) = -\frac{1}{2}\log(2\pi\gamma^2) - \dfrac{(x - w\mu - b)^2 + w^2\sigma^2}{2\gamma^2} - \frac{1}{2}\left(\mu^2 + \sigma^2 - \log\sigma^2 - 1\right)$
8. $\mu^* = \dfrac{w(x - b)}{\gamma^2 + w^2}$, $\sigma^{*2} = \dfrac{\gamma^2}{\gamma^2 + w^2}$; q^* is the posterior $p(z \mid x)$, so the bound is tight: $\mathcal{L}^* = \log p(x) = \log \mathcal{N}(x; b, w^2 + \gamma^2)$
9. $\dfrac{\partial \hat{\mathcal{L}}}{\partial \mu} = g'(z) - \mu$; $\dfrac{\partial \hat{\mathcal{L}}}{\partial \sigma} = g'(z)\,\varepsilon - \sigma + \dfrac{1}{\sigma}$; $\dfrac{\partial \hat{\mathcal{L}}}{\partial s} = \dfrac{\sigma}{2}\, g'(z)\,\varepsilon - \frac{1}{2}(\sigma^2 - 1)$, with $z = \mu + \sigma\varepsilon$; each is an unbiased estimate of the gradient of $\mathcal{L}$
10. $\hat{K} = -\log\sigma + \frac{1}{2}\mu^2 + \mu\sigma\varepsilon + \frac{1}{2}(\sigma^2 - 1)\varepsilon^2$; $\mathbb{E}[\hat{K}] = \mathrm{KL}(q \,\|\, p)$; $\mathrm{Var}(\hat{K}) = \mu^2\sigma^2 + \frac{1}{2}(\sigma^2 - 1)^2$

Worked solutions

Problem 1

> Show that $\log p(x) \geq \mathcal{L}(q) = \mathbb{E}_q[\log p(x, z) - \log q(z)]$ for every admissible q, and say when the two are equal.

$p(x) = \int p(x, z)\, dz = \int q(z)\, \dfrac{p(x, z)}{q(z)}\, dz = \mathbb{E}_q\left[\dfrac{p(x, z)}{q(z)}\right]$.	Multiply and divide the integrand by $q(z)$, which is positive wherever $p(x, z)$ is; an integral against q is an expectation under q.
$\log p(x) = \log \mathbb{E}_q[w] \geq \mathbb{E}_q[\log w]$ with $w = p(x, z)/q(z)$.	Jensen's inequality for the concave logarithm.
$\mathbb{E}_q[\log w] = \mathbb{E}_q[\log p(x, z) - \log q(z)] = \mathcal{L}(q)$.	The log of a quotient is a difference.
Equality holds exactly when w is constant under q: $p(x, z) = c\, q(z)$, and integrating over z gives $c = p(x)$, so $q(z) = p(x, z)/p(x) = p(z \mid x)$.	Jensen's equality condition for a strictly concave function; q integrates to 1.
$\log p(x) \geq \mathcal{L}(q) = \mathbb{E}_q[\log p(x, z) - \log q(z)]$, **with equality exactly when** $q(z) = p(z \mid x)$	The bound holds for every q, which is what makes it usable: pick a family that can be sampled and differentiated, and maximize over it. Problem 2 computes the gap exactly, and Mistake 4 is this derivation with Jensen's inequality the wrong way round.

Problem 2

> Show that $\log p(x) - \mathcal{L}(q) = \mathrm{KL}\left(q(z) \,\|\, p(z \mid x)\right)$.

$\log p(x, z) = \log p(z \mid x) + \log p(x)$.	$p(x, z) = p(z \mid x)\, p(x)$, the product rule of the Bayes page.

$\mathcal{L}(q) = \mathbb{E}_q[\log p(z \mid x) + \log p(x) - \log q(z)] = \log p(x) + \mathbb{E}_q[\log p(z \mid x) - \log q(z)].$	Substitute; $\log p(x)$ does not depend on z, and the expectation of a constant is the constant.
$\mathbb{E}_q[\log p(z \mid x) - \log q(z)] = -\mathbb{E}_q[\log q(z) - \log p(z \mid x)] = -\operatorname{KL}\big(q \,\|\, p(z \mid x)\big).$	The definition of KL with the posterior as the second argument.
$\log p(x) = \mathcal{L}(q) + \operatorname{KL}\big(q(z) \,\|\, p(z \mid x)\big)$, **so** $\log p(x) - \mathcal{L}(q) = \operatorname{KL}\big(q \,\|\, p(z \mid x)\big) \geq 0$	$\operatorname{KL} \geq 0$ (the entropy page's Problem 5) recovers Problem 1 with the gap named. Since $\log p(x)$ does not depend on q, raising $\mathcal{L}$ over q is the same as lowering the KL to the true posterior, without ever computing that posterior or its normalizer $p(x)$. Problem 8 shows the gap closing to 0 when the family contains the posterior.

Problem 3

> Show that $\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid z)] - \operatorname{KL}\big(q(z) \,\|\, p(z)\big) = \mathbb{E}_q[\log p(x \mid z)] + \mathbb{E}_q[\log p(z)] + H(q)$.

$\log p(x, z) = \log p(x \mid z) + \log p(z).$	The product rule the other way round.
$\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid z)] + \mathbb{E}_q[\log p(z)] - \mathbb{E}_q[\log q(z)].$	Substitute into the definition and use linearity of $\mathbb{E}_q$.
$\mathbb{E}_q[\log p(z)] - \mathbb{E}_q[\log q(z)] = -\mathbb{E}_q[\log q(z) - \log p(z)] = -\operatorname{KL}\big(q \,\|\, p(z)\big).$	The definition of KL with the prior as the second argument.
$-\mathbb{E}_q[\log q(z)] = H(q).$	The definition of entropy.

$\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid z)] - \mathrm{KL}\left(q(z) \,\|\, p(z)\right) = \mathbb{E}_q[\log p(x \mid z)] + \mathbb{E}_q[\log p(z)] + H(q)$	The first form is the VAE objective: a reconstruction term, the expected log-likelihood of x under the decoder, minus a regularizer that pulls q toward the prior. For Gaussian q and prior the KL is closed-form (Problem 4), so only the reconstruction term needs sampling (Problem 9). The second form shows what the entropy does: without $H(q)$, maximizing $\mathbb{E}_q[\log p(x, z)]$ over q would collapse q onto the single z that maximizes $p(x, z)$, and the bound would no longer be a bound.

Problem 4

Let $q = \mathcal{N}(\mu_1, \mathrm{diag}(\sigma_1^2))$ and $p = \mathcal{N}(\mu_2, \mathrm{diag}(\sigma_2^2))$ on $\mathbb{R}^d$. Show that

$$\mathrm{KL}(q \,\|\, p) = \sum_{j=1}^{d} \Big(\log \frac{\sigma_{2j}}{\sigma_{1j}} + \frac{\sigma_{1j}^2 + (\mu_{1j} - \mu_{2j})^2}{2\sigma_{2j}^2} - \frac{1}{2} \Big),$$

and that with $\mu_2 = 0$, $\sigma_2 = \mathbf{1}$ and $s_j = \log \sigma_{1j}^2$ it becomes $\tfrac{1}{2} \sum_j \left(\mu_{1j}^2 + e^{s_j} - s_j - 1 \right)$.

$q(z) = \prod_j q_j(z_j)$ with $q_j = \mathcal{N}(\mu_{1j}, \sigma_{1j}^2)$, and likewise $p(z) = \prod_j p_j(z_j)$.	A diagonal Gaussian has independent coordinates, so its density is the product of the one-dimensional densities.
$\log q(z) - \log p(z) = \sum_j \left(\log q_j(z_j) - \log p_j(z_j) \right)$.	The log of a product is a sum.
$\mathbb{E}_q\big[\log q_j(z_j) - \log p_j(z_j) \big] = \mathrm{KL}(q_j \,\|\, p_j)$.	The expectation of a function of z_j alone uses only the marginal of z_j under q, which is q_j.
$\mathrm{KL}(q \,\|\, p) = \sum_j \mathrm{KL}(q_j \,\|\, p_j) = \sum_j \Big(\log \frac{\sigma_{2j}}{\sigma_{1j}} + \frac{\sigma_{1j}^2 + (\mu_{1j} - \mu_{2j})^2}{2\sigma_{2j}^2} - \frac{1}{2} \Big)$.	Linearity of $\mathbb{E}_q$ over step 2, then the entropy page's one-dimensional formula for each pair.

With $\mu_{2j}=0$, $\sigma_{2j}=1$ and $\sigma_{1j}^2=e^{s_j}$: $\log\dfrac{1}{\sigma_{1j}}=-\tfrac{1}{2}s_j$ and $\dfrac{\sigma_{1j}^2+\mu_{1j}^2}{2}-\dfrac{1}{2}=\tfrac{1}{2}\left(e^{s_j}+\mu_{1j}^2-1\right)$.	$\log\sigma_{1j}=\tfrac{1}{2}\log\sigma_{1j}^2=\tfrac{1}{2}s_j$.
$\mathrm{KL}(q\,\|\,p)=\sum_j\left(\log\dfrac{\sigma_{2j}}{\sigma_{1j}}+\dfrac{\sigma_{1j}^2+(\mu_{1j}-\mu_{2j})^2}{2\sigma_{2j}^2}-\dfrac{1}{2}\right)$; **against** $\mathcal{N}(0,I)$ **with** $s_j=\log\sigma_{1j}^2$ **it is** $\tfrac{1}{2}\sum_j\left(\mu_{1j}^2+e^{s_j}-s_j-1\right)$	The KL between two product distributions is the sum of the coordinate KLs, for any product distributions, not only Gaussians. The check computes each coordinate's expectation by quadrature and also the full multivariate formula $\tfrac{1}{2}\big[\,\mathrm{tr}(\Sigma_2^{-1}\Sigma_1)+(\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1)-d+\log\det\Sigma_2-\log\det\Sigma_1\big]$ with diagonal Σ's; all three agree. The $\log\sigma^2$ parameterization keeps $\sigma_j>0$ without a constraint; the gradient in s is on the entropy page (Problem 9) and reappears in Problem 9 here.

Problem 5

> Let $z=\mu+\sigma\odot\varepsilon$ with $\varepsilon\sim\mathcal{N}(0,I)$, and $F(\mu,\sigma)=\mathbb{E}[f(z)]$ for a smooth $f:\mathbb{R}^d\to\mathbb{R}$. Show that $\nabla_\mu F=\mathbb{E}[\nabla f(z)]$, that $\partial F/\partial\sigma_j=\mathbb{E}[\partial_j f(z)\,\varepsilon_j]$, and that with $s_j=\log\sigma_j^2$, $\partial F/\partial s_j=\tfrac{\sigma_j}{2}\,\mathbb{E}[\partial_j f(z)\,\varepsilon_j]$. Verify all three on $f(z)=z^\top Bz+c^\top z$ with B symmetric, where $F=\mu^\top B\mu+\sum_j B_{jj}\sigma_j^2+c^\top\mu$.

$F(\mu,\sigma)=\mathbb{E}_\varepsilon\big[f(\mu+\sigma\odot\varepsilon)\big]$, where the distribution of ε involves neither μ nor σ.	The reparameterization: the parameters have moved from the measure into the integrand, which is what allows the next step.
$\nabla_\mu F=\mathbb{E}_\varepsilon\big[\nabla_\mu f(\mu+\sigma\odot\varepsilon)\big]=\mathbb{E}[\nabla f(z)]$.	Differentiate under the expectation, which is legitimate for a smooth f whose gradient has finite expectation; $\partial z/\partial\mu=I$.
$\dfrac{\partial F}{\partial\sigma_j}=\mathbb{E}\Big[\partial_j f(z)\,\dfrac{\partial z_j}{\partial\sigma_j}\Big]=\mathbb{E}[\partial_j f(z)\,\varepsilon_j]$.	Only z_j depends on σ_j, with $\partial z_j/\partial\sigma_j=\varepsilon_j$; chain rule.

$\sigma_j = e^{s_j/2}$ gives $\partial\sigma_j/\partial s_j = \frac{1}{2}e^{s_j/2} = \frac{\sigma_j}{2}$, so $\dfrac{\partial F}{\partial s_j} = \dfrac{\sigma_j}{2}\,\mathbb{E}[\partial_j f(z)\,\varepsilon_j]$.	Chain rule through $\sigma_j(s_j)$.
For the quadratic, $\mathbb{E}[zz^\top] = \mu\mu^\top + \operatorname{diag}(\sigma^2)$, so $F = \operatorname{tr}\left(B\,\mathbb{E}[zz^\top]\right) + c^\top\mu = \mu^\top B\mu + \sum_j B_{jj}\sigma_j^2 + c^\top\mu$.	$z^\top Bz = \operatorname{tr}(Bzz^\top)$ and $\mathbb{E}$ is linear; the variance page's $\mathbb{E}[zz^\top] = \Sigma + \mu\mu^\top$ with $\Sigma = \operatorname{diag}(\sigma^2)$; $\operatorname{tr}(B\operatorname{diag}(\sigma^2)) = \sum_j B_{jj}\sigma_j^2$.
Differentiating F itself gives $\nabla_\mu F = 2B\mu + c$, $\partial F/\partial\sigma_j = 2B_{jj}\sigma_j$ and $\partial F/\partial s_j = B_{jj}\sigma_j^2$. The estimators: $\nabla f(z) = 2Bz + c$ has expectation $2B\mu + c$, and $\partial_j f(z)\,\varepsilon_j = \big(2\sum_k B_{jk}(\mu_k + \sigma_k\varepsilon_k) + c_j\big)\varepsilon_j$ has expectation $2B_{jj}\sigma_j$.	The matrix-calculus page for $\nabla_\mu(\mu^\top B\mu) = 2B\mu$; $\mathbb{E}[\varepsilon_j] = 0$ kills the μ and c terms, and $\mathbb{E}[\varepsilon_k\varepsilon_j] = 1$ for $k = j$ and 0 otherwise keeps one term of the sum.
$\nabla_\mu F = \mathbb{E}[\nabla f(z)]$, $\dfrac{\partial F}{\partial\sigma_j} = \mathbb{E}[\partial_j f(z)\,\varepsilon_j]$, $\dfrac{\partial F}{\partial s_j} = \dfrac{\sigma_j}{2}\,\mathbb{E}[\partial_j f(z)\,\varepsilon_j]$; **for the quadratic both routes give** $2B\mu + c$, $2B_{jj}\sigma_j$ **and** $B_{jj}\sigma_j^2$	This is the reparameterization gradient: draw ε, form z, backpropagate $\nabla f(z)$ through $z = \mu + \sigma \odot \varepsilon$, so that μ receives $\nabla f(z)$ and σ receives $\nabla f(z) \odot \varepsilon$. One draw gives an unbiased estimate of each gradient; the check averages 400,000 draws and matches to Monte Carlo accuracy. In a VAE, f is the decoder's $\log p_\theta(x \mid z)$, and the alternative that differentiates the density instead (Problem 6) is far noisier.

Problem 6

> Take $q = \mathcal{N}(\mu, \sigma^2)$ in one dimension and $f(z) = z^2$. Two unbiased estimators of $\partial\,\mathbb{E}_q[f(z)]/\partial\mu$ are the reparameterized $g_R = f'(\mu + \sigma\varepsilon) = 2(\mu + \sigma\varepsilon)$ and the score-function $g_S = f(z)\,\partial_\mu \log q(z) = z^2(z - \mu)/\sigma^2$ with $z = \mu + \sigma\varepsilon$. Show both have mean 2μ and compute their variances.

$\dfrac{\partial}{\partial\mu}\mathbb{E}_q[z^2] = \dfrac{\partial}{\partial\mu}(\mu^2 + \sigma^2) = 2\mu.$	$\mathbb{E}[z^2] = \operatorname{Var}(z) + (\mathbb{E}[z])^2$, the variance page's Problem 1.

$\mathbb{E}[g_R] = 2\mu + 2\sigma\,\mathbb{E}[\varepsilon] = 2\mu$ and $\mathrm{Var}(g_R) = 4\sigma^2\,\mathrm{Var}(\varepsilon) = 4\sigma^2$.	$\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = 1$; the variance page's Problem 1 for the scaling.
$g_S = \dfrac{(\mu + \sigma\varepsilon)^2\,\varepsilon}{\sigma} = \dfrac{\mu^2\varepsilon + 2\mu\sigma\varepsilon^2 + \sigma^2\varepsilon^3}{\sigma}$.	$(z-\mu)/\sigma^2 = \varepsilon/\sigma$; expand the square.
$\mathbb{E}[g_S] = \dfrac{\mu^2\cdot 0 + 2\mu\sigma\cdot 1 + \sigma^2\cdot 0}{\sigma} = 2\mu.$	$\mathbb{E}[\varepsilon] = \mathbb{E}[\varepsilon^3] = 0$ and $\mathbb{E}[\varepsilon^2] = 1$.
$\mathbb{E}[g_S^2] = \dfrac{\mathbb{E}[(\mu^2\varepsilon + 2\mu\sigma\varepsilon^2 + \sigma^2\varepsilon^3)^2]}{\sigma^2} = \dfrac{\mu^4 + 12\mu^2\sigma^2 + 15\sigma^4 + 6\mu^2\sigma^2}{\sigma^2} = \dfrac{\mu^4}{\sigma^2} + 18\mu^2 + 15\sigma^2$.	Square the three-term sum: the squares give $\mu^4\mathbb{E}[\varepsilon^2]$, $4\mu^2\sigma^2\mathbb{E}[\varepsilon^4]$ and $\sigma^4\mathbb{E}[\varepsilon^6]$; of the cross terms only $2\mu^2\sigma^2\mathbb{E}[\varepsilon^4]$ survives, the others carrying odd powers of ε; then $\mathbb{E}[\varepsilon^4] = 3$ and $\mathbb{E}[\varepsilon^6] = 15$.
$\mathrm{Var}(g_S) = \dfrac{\mu^4}{\sigma^2} + 18\mu^2 + 15\sigma^2 - 4\mu^2 = \dfrac{\mu^4}{\sigma^2} + 14\mu^2 + 15\sigma^2$.	Subtract the squared mean.
Both estimators have mean 2μ; $\mathrm{Var}(g_R) = 4\sigma^2$ **while** $\mathrm{Var}(g_S) = \dfrac{\mu^4}{\sigma^2} + 14\mu^2 + 15\sigma^2$	At $\mu = \sigma = 1$ the variances are 4 and 30; as $\sigma \to 0$ the score-function variance grows like μ^4/σ^2 while the gradient it estimates stays at 2μ. The reparameterized estimator uses f', so each sample reports which way to move; the score-function estimator uses only values of f and must infer the direction from which samples scored higher. This gap is why VAEs reparameterize and why policy gradients, which cannot (the next page), need baselines.

Problem 7

A linear Gaussian model: $p(z) = \mathcal{N}(0, 1)$, $p(x \mid z) = \mathcal{N}(wz + b, \gamma^2)$ with w, b, γ fixed, and $q(z) = \mathcal{N}(\mu, \sigma^2)$. Show that

$$\mathcal{L}(\mu, \sigma^2) = -\tfrac{1}{2}\log(2\pi\gamma^2) - \frac{(x - w\mu - b)^2 + w^2\sigma^2}{2\gamma^2} - \tfrac{1}{2}\big(\mu^2 + \sigma^2 - \log\sigma^2 - 1\big).$$

$\log p(x \mid z) = -\tfrac{1}{2}\log(2\pi\gamma^2) - \dfrac{(x - wz - b)^2}{2\gamma^2}.$	The Gaussian log-density, the multivariate-Gaussian page's Problem 1 with $d = 1$.
$\mathbb{E}_q[(x - wz - b)^2] = \mathrm{Var}_q(x - wz - b) + \big(\mathbb{E}_q[x - wz - b]\big)^2 = w^2\sigma^2 + (x - w\mu - b)^2.$	The variance page's Problem 1: $\mathbb{E}[u^2] = \mathrm{Var}(u) + (\mathbb{E}[u])^2$, with $\mathrm{Var}(-wz) = w^2\sigma^2$ and $\mathbb{E}_q[z] = \mu$.
$\mathbb{E}_q[\log p(x \mid z)] = -\tfrac{1}{2}\log(2\pi\gamma^2) - \dfrac{(x - w\mu - b)^2 + w^2\sigma^2}{2\gamma^2}.$	Linearity of $\mathbb{E}_q$ over step 1.
$\mathrm{KL}\left(q \,\|\, p(z)\right) = \tfrac{1}{2}\big(\mu^2 + \sigma^2 - \log\sigma^2 - 1\big).$	Problem 4 with $d = 1$, $\mu_2 = 0$, $\sigma_2 = 1$.
$\mathcal{L}(\mu, \sigma^2) = -\tfrac{1}{2}\log(2\pi\gamma^2) - \dfrac{(x - w\mu - b)^2 + w^2\sigma^2}{2\gamma^2} - \tfrac{1}{2}\big(\mu^2 + \sigma^2 - \log\sigma^2 - 1\big)$	Problem 3's first form, term by term. The reconstruction term penalizes the mean error and also the spread, through $w^2\sigma^2$: a wide q decodes to a wide range of x. The KL penalizes a narrow q through $-\log\sigma^2$. The two pull in opposite directions and Problem 8 finds the balance; because everything is closed-form here, the model is a test case for the sampled gradients of Problem 9.

Problem 8

> Maximize Problem 7's $\mathcal{L}$ over μ and σ^2. Show that $\mu^* = \dfrac{w(x-b)}{\gamma^2+w^2}$ and $\sigma^{*2} = \dfrac{\gamma^2}{\gamma^2+w^2}$, that $q^* = \mathcal{N}(\mu^*, \sigma^{*2})$ is the exact posterior $p(z \mid x)$, and that the bound is tight: $\mathcal{L}^* = \log p(x)$ with $p(x) = \mathcal{N}(x; b, w^2+\gamma^2)$.

$\dfrac{\partial \mathcal{L}}{\partial \mu} = \dfrac{w(x - w\mu - b)}{\gamma^2} - \mu = 0$ gives $\mu\left(1 + \dfrac{w^2}{\gamma^2}\right) = \dfrac{w(x-b)}{\gamma^2}$, so $\mu^* = \dfrac{w(x-b)}{\gamma^2+w^2}$.	Chain rule on the square, whose inner derivative in μ is $-w$; the KL contributes $-\mu$; multiply through by γ^2.
$\dfrac{\partial \mathcal{L}}{\partial \sigma^2} = -\dfrac{w^2}{2\gamma^2} - \dfrac{1}{2} + \dfrac{1}{2\sigma^2} = 0$ gives $\dfrac{1}{\sigma^2} = 1 + \dfrac{w^2}{\gamma^2}$, so $\sigma^{*2} = \dfrac{\gamma^2}{\gamma^2+w^2}$.	Differentiate with σ^2 as the variable: $-\log \sigma^2$ has derivative $-1/\sigma^2$.
The stationary point is a maximum: $\partial^2\mathcal{L}/\partial\mu^2 = -w^2/\gamma^2 - 1 < 0$, $\partial^2\mathcal{L}/\partial(\sigma^2)^2 = -1/(2\sigma^4) < 0$ and the mixed derivative is 0.	A diagonal Hessian with negative entries is negative definite, the Hessians page's test.
The exact posterior is $p(z \mid x) \propto p(z)\,p(x \mid z) \propto \exp\left(-\dfrac{z^2}{2} - \dfrac{(x - wz - b)^2}{2\gamma^2}\right)$, a Gaussian in z with precision $1 + w^2/\gamma^2$ and mean $\dfrac{w(x-b)/\gamma^2}{1 + w^2/\gamma^2} = \dfrac{w(x-b)}{\gamma^2+w^2}$: exactly (μ^*, σ^{*2}).	Collect the z^2 and z terms in the exponent: the coefficient of $-z^2/2$ is the precision and the coefficient of z divided by the precision is the mean, as in the multivariate-Gaussian page's product of Gaussians.
$q^* = p(z \mid x)$ makes the KL of Problem 2 zero, so $\mathcal{L}^* = \log p(x)$; and $x = wz + b + \eta$ with $z \sim \mathcal{N}(0,1)$ and $\eta \sim \mathcal{N}(0, \gamma^2)$ independent has mean b and variance $w^2 + \gamma^2$.	Problem 2; the multivariate-Gaussian page's affine image and the variance page's Problem 2 for the sum of independent terms.

$\mu^* = \dfrac{w(x-b)}{\gamma^2+w^2}$, $\sigma^{*2} = \dfrac{\gamma^2}{\gamma^2+w^2}$; q^* **is the posterior** $p(z \mid x)$**, so the bound is tight:** $\mathcal{L}^* = \log p(x) = \log \mathcal{N}(x; b, w^2+\gamma^2)$	The variational family contains the posterior, so the ELBO reaches the evidence; with a nonlinear decoder it does not, and the gap of Problem 2 remains. The posterior is narrower than the prior, $\sigma^{*2} < 1$, by more when the signal-to-noise ratio w^2/γ^2 is large, and μ^* shrinks the naive estimate $(x-b)/w$ toward 0 by the factor $w^2/(\gamma^2+w^2)$. A VAE's encoder learns an amortized version of the map $x \mapsto (\mu^*, \sigma^*)$.

Problem 9

Let the decoder give $\log p_\theta(x \mid z) = g(z)$ for a differentiable g, with $q = \mathcal{N}(\mu, \sigma^2)$ and prior $\mathcal{N}(0,1)$. The one-sample estimate of the ELBO is $\hat{\mathcal{L}}(\mu,\sigma) = g(\mu+\sigma\varepsilon) - \frac{1}{2}\big(\mu^2+\sigma^2-\log\sigma^2-1\big)$ with ε drawn once. Compute $\partial\hat{\mathcal{L}}/\partial\mu$, $\partial\hat{\mathcal{L}}/\partial\sigma$ and $\partial\hat{\mathcal{L}}/\partial s$ with $s = \log\sigma^2$, and show that each has expectation equal to the corresponding gradient of $\mathcal{L}$.

$z = \mu + \sigma\varepsilon$ with $\partial z/\partial\mu = 1$ and $\partial z/\partial\sigma = \varepsilon$.	The reparameterization; ε is a constant once drawn.
$\dfrac{\partial\hat{\mathcal{L}}}{\partial\mu} = g'(z) - \mu.$	Chain rule through z; the KL term's derivative in μ is μ.
$\dfrac{\partial\hat{\mathcal{L}}}{\partial\sigma} = g'(z)\,\varepsilon - \sigma + \dfrac{1}{\sigma}.$	Chain rule through z; $\frac{1}{2}(\sigma^2 - \log\sigma^2)$ has derivative $\sigma - 1/\sigma$.
$\dfrac{\partial\hat{\mathcal{L}}}{\partial s} = \dfrac{\sigma}{2}\,g'(z)\,\varepsilon - \tfrac{1}{2}(\sigma^2-1).$	$\sigma = e^{s/2}$ gives $\partial\sigma/\partial s = \sigma/2$ for the first term; the KL in s is $\frac{1}{2}(\mu^2+e^s-s-1)$ with derivative $\frac{1}{2}(e^s-1)$, the entropy page's Problem 9.
$\mathbb{E}_\varepsilon\big[\partial\hat{\mathcal{L}}/\partial\mu\big] = \mathbb{E}[g'(z)] - \mu = \partial_\mu\mathbb{E}[g(z)] - \partial_\mu \mathrm{KL}$, and likewise for σ and s.	Problem 5 with $f = g$ for the reconstruction term; the KL term contains no ε and is already exact.

$\frac{\partial \hat{\mathcal{L}}}{\partial \mu} = g'(z) - \mu$; $\frac{\partial \hat{\mathcal{L}}}{\partial \sigma} = g'(z)\,\varepsilon - \sigma + \frac{1}{\sigma}$; $\frac{\partial \hat{\mathcal{L}}}{\partial s} = \frac{\sigma}{2}\,g'(z)\,\varepsilon - \frac{1}{2}(\sigma^2 - 1)$, **with** $z = \mu + \sigma\varepsilon$; **each is an unbiased estimate of the gradient of** $\mathcal{L}$	This is one VAE training step for one latent coordinate: the decoder's gradient $g'(z)$ at the sampled z, passed back to μ with weight 1 and to σ with weight ε, plus the closed-form KL gradient. Treating the sample as a constant (Mistake 2) deletes $g'(z)\varepsilon$ and the encoder's variance stops learning from the data. With d coordinates everything is per coordinate by Problem 4, and the check differentiates $\hat{\mathcal{L}}$ numerically for a specific nonlinear g with ε held fixed.

Problem 10

> Instead of Problem 4's closed form, the KL can be estimated from one sample: $\hat{K} = \log q(z) - \log p(z)$ with $z = \mu + \sigma\varepsilon$, $q = \mathcal{N}(\mu, \sigma^2)$ and $p = \mathcal{N}(0, 1)$. Show that $\hat{K} = -\log\sigma + \frac{1}{2}\mu^2 + \mu\sigma\varepsilon + \frac{1}{2}(\sigma^2 - 1)\varepsilon^2$, that $\mathbb{E}[\hat{K}] = \mathrm{KL}(q \,\|\, p)$, and that $\mathrm{Var}(\hat{K}) = \mu^2\sigma^2 + \frac{1}{2}(\sigma^2 - 1)^2$.

$\log q(z) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{(z-\mu)^2}{2\sigma^2} = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2}\varepsilon^2.$	$(z - \mu)/\sigma = \varepsilon$.
$\log p(z) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(\mu + \sigma\varepsilon)^2 = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\mu^2 - \mu\sigma\varepsilon - \frac{1}{2}\sigma^2\varepsilon^2.$	Expand the square.
$\hat{K} = -\log\sigma - \frac{1}{2}\varepsilon^2 + \frac{1}{2}\mu^2 + \mu\sigma\varepsilon + \frac{1}{2}\sigma^2\varepsilon^2 = -\log\sigma + \frac{1}{2}\mu^2 + \mu\sigma\varepsilon + \frac{1}{2}(\sigma^2 - 1)\varepsilon^2.$	Subtract; the $\log(2\pi)$ terms cancel and the ε^2 terms combine.
$\mathbb{E}[\hat{K}] = -\log\sigma + \frac{1}{2}\mu^2 + 0 + \frac{1}{2}(\sigma^2 - 1) = \frac{1}{2}\big(\mu^2 + \sigma^2 - \log\sigma^2 - 1\big).$	$\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon^2] = 1$, and $\log\sigma = \frac{1}{2}\log\sigma^2$; this is Problem 4's closed form.
$\mathrm{Var}(\hat{K}) = \mu^2\sigma^2\,\mathrm{Var}(\varepsilon) + \frac{1}{4}(\sigma^2 - 1)^2\,\mathrm{Var}(\varepsilon^2) + 2\cdot\mu\sigma\cdot\frac{1}{2}(\sigma^2 - 1)\,\mathrm{Cov}(\varepsilon, \varepsilon^2) = \mu^2\sigma^2 + \frac{1}{2}(\sigma^2 - 1)^2.$	The variance page's Problem 2 on the two random terms of step 3; $\mathrm{Var}(\varepsilon^2) = \mathbb{E}[\varepsilon^4] - 1 = 2$ and $\mathrm{Cov}(\varepsilon, \varepsilon^2) = \mathbb{E}[\varepsilon^3] - \mathbb{E}[\varepsilon]\,\mathbb{E}[\varepsilon^2] = 0$.

$\hat{K} = -\log\sigma + \frac{1}{2}\mu^2 + \mu\sigma\varepsilon + \frac{1}{2}(\sigma^2 - 1)\varepsilon^2$; $\mathbb{E}[\hat{K}] = \mathrm{KL}(q \,\|\, p)$; $\mathrm{Var}(\hat{K}) = \mu^2\sigma^2 + \frac{1}{2}(\sigma^2 - 1)^2$	The estimator is unbiased, so a VAE trained with it is right on average, but its noise is pure cost: at $\mu = 2$, $\sigma = 1$ the variance is 4 for a KL of 2. The closed form removes it entirely, which is why implementations compute the KL analytically whenever q and the prior are Gaussian and sample only the reconstruction term. At $q = p$ the estimator is identically 0, so near the prior it is quiet; sampling from the prior instead of q (Mistake 3) breaks even the unbiasedness.

Where this goes wrong

1. The KL term written the other way round

KL is not symmetric, and the regularizer is often described as the distance between q and the prior.

$\mathcal{L} = \mathbb{E}_q[\log p(x \mid z)] - \mathrm{KL}\left(q(z) \,\|\, p(z)\right)$	Right so far: Problem 3.
"KL is the distance from q to the prior, so the order does not matter."	The habit that causes the mistake: treating KL as a symmetric distance, which it is not (the entropy page's Problem 6).
$\mathcal{L} = \mathbb{E}_q[\log p(x \mid z)] - \mathrm{KL}\left(p(z) \,\|\, q(z)\right)$ ✗	For $q = \mathcal{N}(\mu, \sigma^2)$ and $p = \mathcal{N}(0, 1)$ the reversed term is $\log\sigma + \dfrac{1 + \mu^2}{2\sigma^2} - \dfrac{1}{2}$, which at $\mu = 0$, $\sigma = 2$ is 0.318 where the correct $\frac{1}{2}(\sigma^2 - \log\sigma^2 - 1)$ is 0.807. The result is no longer a lower bound on $\log p(x)$ and Problem 2's identity fails. Its gradient in σ is different in kind: the reversed KL punishes a small σ through $1/\sigma^2$ and a large one only logarithmically, so it pushes q to be wide, the mode-covering behavior of the entropy page's Problem 7 rather than the mode-seeking one the ELBO has.

2. Treating the sample as a constant

After sampling, z is a tensor of numbers, and `dist.sample()` returns exactly that, with no gradient path back to μ or σ.

$\hat{\mathcal{L}} = g(z) - \frac{1}{2}(\mu^2 + \sigma^2 - \log\sigma^2 - 1)$ with $z = \mu + \sigma\varepsilon$	Right so far: Problem 9.
"z has been sampled, so it is a number now; only the KL term depends on σ."	The shortcut that causes the mistake: sampling with `sample()` instead of `rsample()`, which detaches z from its parameters.
$\dfrac{\partial\hat{\mathcal{L}}}{\partial\sigma} = -\sigma + \dfrac{1}{\sigma}$ ✗	The reconstruction term's dependence on σ through $z = \mu + \sigma\varepsilon$ is gone, and with it $g'(z)\varepsilon$ (Problem 9); likewise $\partial\hat{\mathcal{L}}/\partial\mu$ loses $g'(z)$. What remains is the KL's gradient alone, which is zero at $\sigma = 1$, $\mu = 0$ and pushes the encoder there whatever the data say: q collapses onto the prior and the decoder is trained on noise. The reparameterization of Problem 5 exists to keep that path; z must be written as a function of (μ, σ) and ε before anything is differentiated.

3. Monte Carlo KL averaged over samples from the prior

The KL is an integral of $\log q - \log p$, and the prior $\mathcal{N}(0, I)$ is the easiest distribution in sight to sample.

$\mathrm{KL}(q \,\|\, p) = \mathbb{E}_q[\log q(z) - \log p(z)]$	Right so far: the definition.
"Draw z from the prior, which needs no parameters, and average $\log q - \log p$."	The shortcut that causes the mistake: an expectation under q estimated with samples from p, which is a different integral.

$\mathrm{KL}(q \,\|\, p) \approx \frac{1}{K}\sum_{k=1}^{K}\big(\log q(z_k) - \log p(z_k)\big)$ with $z_k \sim p$ ✗	Its expectation is $\mathbb{E}_p[\log q - \log p] = -\mathrm{KL}(p \,\|\, q)$, which is never positive, so the estimate is negative on average although a KL is never negative: at $\mu = 0$, $\sigma = 2$ it averages -0.318 where the KL is 0.807. The estimator of Problem 10 draws $z = \mu + \sigma\varepsilon$ from q and is unbiased; the samples must come from the distribution the expectation is under.

4. Jensen's inequality applied in the wrong direction

The inequality says that the log of an average and the average of the logs differ, and which is larger is easy to misremember.

$\log p(x) = \log \mathbb{E}_q\big[p(x,z)/q(z)\big]$	Right so far: Problem 1, step 1.
"Move the log inside: $\log \mathbb{E}[w] \le \mathbb{E}[\log w]$. "	The habit that causes the mistake: the direction for a convex function such as the square, $\mathbb{E}[w^2] \ge (\mathbb{E}[w])^2$, applied to the concave logarithm.
$\log p(x) \le \mathbb{E}_q\big[\log p(x,z) - \log q(z)\big]$ ✗	For the concave logarithm, $\log \mathbb{E}[w] \ge \mathbb{E}[\log w]$, so the ELBO is a lower bound. The check is Problem 2: $\log p(x) - \mathcal{L}(q) = \mathrm{KL}\big(q \,\|\, p(z \mid x)\big) \ge 0$, and on the discrete model in the check $\mathcal{L}$ is strictly below $\log p(x)$ for every q that is not the posterior. Were the line true, maximizing $\mathcal{L}$ would raise an upper bound, which says nothing about $\log p(x)$.

5. Reparameterizing with the variance in place of the standard deviation

The encoder outputs $s = \log \sigma^2$, and the sampling line needs a scale.

$z = \mu + \sigma \odot \varepsilon$ with $\sigma_j = e^{s_j/2}$	Right so far: Problem 5's reparameterization.

“The encoder's scale is e^s, so multiply ε by it.”	The shortcut that causes the mistake: exponentiating the log-variance gives the variance, and the sample needs the standard deviation.
$z = \mu + e^s \odot \varepsilon$ ✗	Then $z \sim \mathcal{N}\big(\mu, \operatorname{diag}(e^{2s})\big) = \mathcal{N}\big(\mu, \operatorname{diag}(\sigma^4)\big)$: the sample's standard deviation is σ^2, too wide when $\sigma > 1$ and too narrow when $\sigma < 1$. The KL term, computed from s for $\mathcal{N}(\mu, \sigma^2)$ (Problem 4), then regularizes a different distribution from the one that was sampled, and the two halves of the ELBO describe two different q's. The scale is $e^{s/2}$, `exp(0.5 * logvar)` in code; the multivariate-Gaussian page's $x = \mu + \Sigma\varepsilon$ is the full-covariance form of the same error.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 24 of *Deep Learning Math by Hand*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/elbo-and-the-reparameterization-trick