Sampling Methods: From MCMC to Generative Modeling
Bayesian learning and Langevin algorithm

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Outline

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^{p}$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^p$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Step 1. Compute the Likelihood:

$$p(\mathcal{D}|x) \overset{(1)}{\propto} \prod_{i=1}^{p} p(y_i|x, w_i) \overset{(2)}{\propto} \exp(-\frac{1}{2}\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp(-\frac{\|x\|^2}{2}).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Denoting $\pi := p(\cdot|\mathcal{D})$ the posterior on parameters $x \in \mathbb{R}^d$, we have:

$$\boxed{\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \frac{1}{2}\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2 + \frac{\|x\|^2}{2}.}$$

i.e. π's density is known "up to a normalization constant".
π is a probability distribution over parameters of a model.

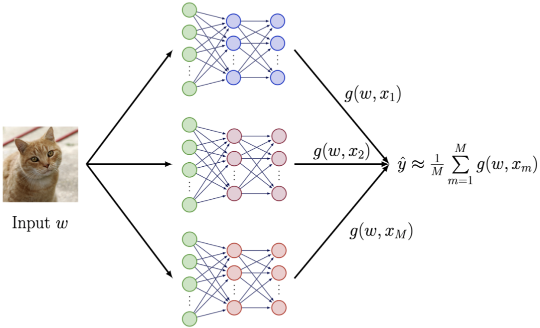
The posterior π is interesting for

- measuring uncertainty on prediction through the distribution of $g(w, \cdot)$, $x \sim \pi$.
- prediction for a new input w:

$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w, x) d\pi(x)}_{\text{"Bayesian model averaging"}}$$

i.e. predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\pi(x)$.

Here, Sampling methods construct an approximation $\mu_M = \frac{1}{M}\sum_{m=1}^{M}\delta_{x_m}$ of π.



Sampling as Optimization

Actually, in many cases (e.g. it is underlying many algorithms), the sampling problem (approximating π) can be viewed as optimization over $\mathcal{P}(\mathbb{R}^d)$:

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

where D is a divergence or distance, hence that is minimized for $\mu = \pi$.

The Kullback-Leibler divergence

D could be the (reverse) Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

We recognize a f-divergence $\int f\left(\frac{\mu}{\pi}\right) d\pi$ where $f(x) = x\log(x)$. Taking $f(x) = -\log(x)$ yields the (forward) KL i.e. $\mathrm{KL}(\pi|\mu)$.

The Kullback-Leibler divergence

D could be the (reverse) Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

We recognize a f-divergence $\int f\left(\frac{\mu}{\pi}\right) d\pi$ where $f(x) = x\log(x)$. Taking $f(x) = -\log(x)$ yields the (forward) KL i.e. $\mathrm{KL}(\pi|\mu)$.

The (reverse) KL as an objective is convenient when the unnormalized density of π is known since it **does not depend on the normalization constant!**

Indeed writing $\pi(x) = e^{-V(x)}/Z$ we have:

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{e^{-V}}(x)\right) d\mu(x) + \log(Z).$$

But, it is not convenient when μ or π are discrete, because the KL is $+\infty$ unless $supp(\mu) \subset supp(\pi)$.

Examples

- (Parametric methods) Variational Inference : Restrict the search space to a parametric families $\{\mu_\theta,\ \theta \in \mathbb{R}^p\}$. The problem rewrites as a finite-dimensional optimization problem (i.e. over $\mathbb{R}^p$):

$$\min_{\theta \in \mathbb{R}^p} \mathrm{D}(\mu_\theta | \pi)$$

- Example: Gaussians with diagonal covariance matrices can be parametrized by $\theta = (m, \sigma) \in \mathbb{R}^{2d}$ (see Bayes by Backprop in the last section)
- Example: use normalizing flows to construct a family $\mu_\theta = f_{\theta\#}p$ and optimize the previous objective[1]. [1] Rezende, D., Mohamed, S. (2015, June). Variational inference with normalizing flows. In International conference on machine learning.

Examples

- (Parametric methods) Variational Inference : Restrict the search space to a parametric families $\{\mu_\theta,\ \theta \in \mathbb{R}^p\}$. The problem rewrites as a finite-dimensional optimization problem (i.e. over $\mathbb{R}^p$):

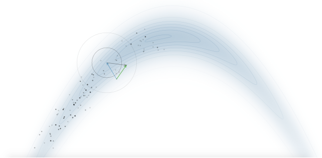
$$\min_{\theta \in \mathbb{R}^p} \mathrm{D}(\mu_\theta | \pi)$$

- Example: Gaussians with diagonal covariance matrices can be parametrized by $\theta = (m, \sigma) \in \mathbb{R}^{2d}$ (see Bayes by Backprop in the last section)
- Example: use normalizing flows to construct a family $\mu_\theta = f_{\theta\#}p$ and optimize the previous objective[1]. [1]Rezende, D., Mohamed, S. (2015, June). Variational inference with normalizing flows. In International conference on machine learning.
- (Non parametric methods) Markov Chain Monte Carlo (MCMC) methods, Sequential Monte Carlo (SMC)...: generate a Markov chain in $\mathbb{R}^d$ whose law converges to $\pi \propto \exp(-V)$
- Example: Langevin (next section)

Langevin Monte Carlo

Langevin Monte Carlo (LMC) [Roberts and Tweedie (1996)]

$$x_{m+1} = x_m + \gamma \nabla \log \pi(x_m) + \sqrt{2\gamma}\eta_m, \quad \eta_m \sim \mathcal{N}(0, \mathrm{Id}).$$



Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

Note that in the Bayesian inference setting, where $\pi = \frac{\exp(-V)}{Z}$, it is easily implementable since the **score** $\nabla_x \log \pi(x) = -\nabla_x(V(x) + \log(Z)) = -\nabla V(x)$ since $\nabla_x \log(Z) = 0$.

Outline

Langevin diffusion

Langevin diffusion is the Stochastic Differential Equation (SDE):

$$\mathrm{d}x_t = -\nabla V(x_t)dt + \sqrt{2}\mathrm{d}B_t, \quad x_t \sim p_t$$

where B_t denotes the standard Brownian motion in $\mathbb{R}^d$, defined as:

- $B_0 = 0$ almost surely;
- For any $t_0 < t_1 < \cdots < t_N$, the increments $B_{t_n} - B_{t_{n-1}}$ are independent, $n = 1, 2, \ldots, N$;
- The difference $B_t - B_s$ and B_{t-s} have the same distribution: $\mathcal{N}(0, (t-s)\,\mathrm{Id})$ for $s < t$;
- B_t is continuous almost surely.

Langevin diffusion defines a *Markov process* as follows:

$$x_t = x_0 - \int_0^t \nabla V(x_s)ds + \sqrt{2}B_t,$$

where x_0 is some initialization.

Time-discretization

An Euler-Maruyama time-discretization of Langevin diffusion yields:

$$x_{t+1} = x_t - \gamma \nabla V(x_t) + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id}). \tag{1}$$

Time-discretization

An Euler-Maruyama time-discretization of Langevin diffusion yields:

$$x_{t+1} = x_t - \gamma \nabla V(x_t) + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id}). \tag{1}$$

Proof:

$$\begin{aligned} x_\gamma &\approx x_0 - \int_0^\gamma \nabla V(x_0)\, dt + \sqrt{2\gamma}\, \eta \\ &= x_0 - \left(\int_0^\gamma dt \right) \nabla V(x_0) + \sqrt{2\gamma}\, \eta \\ &= x_0 - \gamma \nabla V(x_0) + \sqrt{2\gamma}\, \eta. \end{aligned}$$

Time-discretization

An Euler-Maruyama time-discretization of Langevin diffusion yields:

$$x_{t+1} = x_t - \gamma \nabla V(x_t) + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id}). \tag{1}$$

Proof:

$$\begin{aligned} x_\gamma &\approx x_0 - \int_0^\gamma \nabla V(x_0)\, dt + \sqrt{2\gamma}\, \eta \\ &= x_0 - \left(\int_0^\gamma dt \right) \nabla V(x_0) + \sqrt{2\gamma}\, \eta \\ &= x_0 - \gamma \nabla V(x_0) + \sqrt{2\gamma}\, \eta. \end{aligned}$$

We can now iterate this approach k times, which gives us a recursion, which can be easily implementable on a computer:

$$x_{k\gamma} \approx x_{(k-1)\gamma} - \gamma \nabla V(x_{(k-1)\gamma}) + \sqrt{2\gamma}\, \eta_k,$$

where $\eta_k \sim \mathcal{N}(0, \mathrm{Id})$ for all k. Dropping the dependency on γ in the indices yields the scheme (1).

Ornstein-Uhlenbeck

Example: $\pi \propto \exp(-\frac{\|x\|^2}{2})$,

Ornstein-Uhlenbeck

Example: $\pi \propto \exp(-\frac{\|x\|^2}{2})$, $\log \pi(x) = -V(x) = -\frac{\|x\|^2}{2}$, $\nabla \log \pi(x) = -x$.

Ornstein-Uhlenbeck

Example: $\pi \propto \exp(-\frac{\|x\|^2}{2})$, $\log \pi(x) = -V(x) = -\frac{\|x\|^2}{2}$, $\nabla \log \pi(x) = -x$.

(continuous time) **Langevin diffusion** $=$ Ornstein-Uhlenbeck process:

$$\mathrm{d}x_t = -x_t + \mathrm{d}B_t.$$

Ornstein-Uhlenbeck

Example: $\pi \propto \exp(-\frac{\|x\|^2}{2})$, $\log \pi(x) = -V(x) = -\frac{\|x\|^2}{2}$, $\nabla \log \pi(x) = -x$.

(continuous time) **Langevin diffusion** = Ornstein-Uhlenbeck process:

$$\mathrm{d}x_t = -x_t + \mathrm{d}B_t.$$

(discrete time) $x_{t+1} = x_t - \gamma x_t + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id})$.

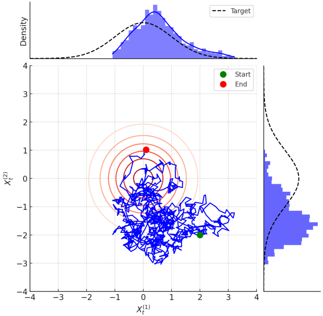
Ornstein-Uhlenbeck

Example: $\pi \propto \exp(-\frac{\|x\|^2}{2})$, $\log \pi(x) = -V(x) = -\frac{\|x\|^2}{2}$, $\nabla \log \pi(x) = -x$.

(continuous time) **Langevin diffusion** = Ornstein-Uhlenbeck process:

$$\mathrm{d}x_t = -x_t + \mathrm{d}B_t.$$

(discrete time) $x_{t+1} = x_t - \gamma x_t + \sqrt{2\gamma}\eta_t, \quad \eta_t \sim \mathcal{N}(0, \mathrm{Id})$.



Recall above we plot $x_{t+1} = x_t + \gamma \nabla \log \pi(x_t) + \sqrt{2\gamma}\eta_t$ for $\pi \propto \exp(-\frac{\|x\|^2}{2})$.

The Fokker-Planck equation

Question: how does the law p_t of x_t evolve? does it converge to π?

The Fokker-Planck equation

Question: how does the law p_t of x_t evolve? does it converge to π?

For simplicity, let us assume $d = 1$, so that Langevin diffusion becomes:

$$\mathrm{d}x_t = -\partial_x V(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

The Fokker-Planck equation

Question: how does the law p_t of x_t evolve? does it converge to π?

For simplicity, let us assume $d = 1$, so that Langevin diffusion becomes:

$$\mathrm{d}x_t = -\partial_x V(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

To understand how $p(x,t)$ evolves, we will use the Fokker–Planck equation, which governs the evolution of $p(x,t)$ through the following partial differential equation (PDE):

$$\partial_t p(x,t) = \partial_x\left[\partial_x V(x) p(x,t)\right] + \partial_x^2 p(x,t).$$

This equation characterizes how the “change” in $p(\cdot,t)$ behaves, i.e., $\partial_t p(x,t)$.

The Fokker-Planck equation

Question: how does the law p_t of x_t evolve? does it converge to π?

For simplicity, let us assume $d = 1$, so that Langevin diffusion becomes:

$$\mathrm{d}x_t = -\partial_x V(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

To understand how $p(x,t)$ evolves, we will use the Fokker–Planck equation, which governs the evolution of $p(x,t)$ through the following partial differential equation (PDE):

$$\partial_t p(x,t) = \partial_x\left[\partial_x V(x) p(x,t)\right] + \partial_x^2 p(x,t).$$

This equation characterizes how the "change" in $p(\cdot,t)$ behaves, i.e., $\partial_t p(x,t)$.

Remark: for $d > 1$, the Fokker-Planck equation writes:

$$\partial_t p(x,t) = \nabla \cdot (\nabla V(x) p(x,t)) + \Delta(p(x,t)).$$

(where $\nabla\cdot$ and Δ are the divergence and Laplacian operators: analog to above but summing all partial derivatives for $x_1, \ldots, x_d$).

The Fokker-Planck equation

Now, the idea is: if $p(\cdot, t)$ converges to a distribution as $t \to \infty$, then whenever this limit is reached, there should not be any more changes in p. In other words, whenever $p(\cdot, t)$ hits its limit, $\partial_t p(x, t)$ has to be equal to 0.

The Fokker-Planck equation

Now, the idea is: if $p(\cdot, t)$ converges to a distribution as $t \to \infty$, then whenever this limit is reached, there should not be any more changes in p. In other words, whenever $p(\cdot, t)$ hits its limit, $\partial_t p(x, t)$ has to be equal to 0.

Therefore, we can simply "check" if $\pi \propto \exp(-V)$ is a limit of $p(\cdot, t)$ by replacing $p(x, t)$ with $\pi(x)$ in the Fokker–Planck equation and observing whether the right-hand side is equal to 0 or not. Let us apply this procedure:

The Fokker-Planck equation

Now, the idea is: if $p(\cdot, t)$ converges to a distribution as $t \to \infty$, then whenever this limit is reached, there should not be any more changes in p. In other words, whenever $p(\cdot, t)$ hits its limit, $\partial_t p(x, t)$ has to be equal to 0.

Therefore, we can simply "check" if $\pi \propto \exp(-V)$ is a limit of $p(\cdot, t)$ by replacing $p(x, t)$ with $\pi(x)$ in the Fokker–Planck equation and observing whether the right-hand side is equal to 0 or not. Let us apply this procedure:

$$\begin{aligned}\partial_x \left[\partial_x V(x)\pi(x)\right] + \partial_x^2 \pi(x) &= \partial_x \left[\partial_x V(x)\pi(x) + \partial_x \pi(x)\right] \\ &= \partial_x \left[\partial_x V(x)\pi(x) - \partial_x V(x)\pi(x)\right] \\ &= 0,\end{aligned}$$

where we used the fact that

$$\partial_x V(x) = -\partial_x \log \pi(x) = -\frac{1}{\pi(x)} \partial_x \pi(x),$$

hence

$$\partial_x \pi(x) = -\pi(x)\partial_x V(x).$$

The Fokker-Planck equation

Now, the idea is: if $p(\cdot, t)$ converges to a distribution as $t \to \infty$, then whenever this limit is reached, there should not be any more changes in p. In other words, whenever $p(\cdot, t)$ hits its limit, $\partial_t p(x, t)$ has to be equal to 0.

Therefore, we can simply "check" if $\pi \propto \exp(-V)$ is a limit of $p(\cdot, t)$ by replacing $p(x, t)$ with $\pi(x)$ in the Fokker–Planck equation and observing whether the right-hand side is equal to 0 or not. Let us apply this procedure:

$$\begin{aligned} \partial_x \left[\partial_x V(x)\pi(x)\right] + \partial_x^2 \pi(x) &= \partial_x \left[\partial_x V(x)\pi(x) + \partial_x \pi(x)\right] \\ &= \partial_x \left[\partial_x V(x)\pi(x) - \partial_x V(x)\pi(x)\right] \\ &= 0, \end{aligned}$$

where we used the fact that

$$\partial_x V(x) = -\partial_x \log \pi(x) = -\frac{1}{\pi(x)}\partial_x \pi(x),$$

hence

$$\partial_x \pi(x) = -\pi(x)\partial_x V(x).$$

Conclusion: π is an equilibrium for the FP equation !

Ornstein–Uhlenbeck Process

We now focus on a specific case of a Langevin diffusion and we will prove that:

For the SDE:

$$dX_t = -\beta X_t\, dt + \sigma\, dB_t$$

The solution is:

$$X_t = e^{-\beta t} X_0 + \sigma e^{-\beta t} \int_0^t e^{\beta s}\, dB_s$$

with stationary/limiting distribution $\pi = \mathcal{N}(0, \frac{\sigma^2}{2\beta})$
and we have:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right)$$

Observe that:

- The farther into the future, the more the initial value gets "forgotten"

Proof

Step 1 (Multiply by the integrating factor)
Multiply both sides of the SDE by $\mu(t) = e^{\beta t}$:

$$e^{\beta t} dX_t = -\beta e^{\beta t} X_t \, dt + \sigma e^{\beta t} dB_t$$

Proof

Step 1 (Multiply by the integrating factor)
Multiply both sides of the SDE by $\mu(t) = e^{\beta t}$:

$$e^{\beta t} dX_t = -\beta e^{\beta t} X_t \, dt + \sigma e^{\beta t} dB_t$$

But using (Itô's) product rule:

$$d\left(e^{\beta t} X_t\right) = e^{\beta t} dX_t + \beta e^{\beta t} X_t \, dt$$

Proof

Step 1 (Multiply by the integrating factor)
Multiply both sides of the SDE by $\mu(t) = e^{\beta t}$:

$$e^{\beta t} dX_t = -\beta e^{\beta t} X_t \, dt + \sigma e^{\beta t} dB_t$$

But using (Itô's) product rule:

$$d\left(e^{\beta t} X_t\right) = e^{\beta t} dX_t + \beta e^{\beta t} X_t \, dt$$

So we get:

$$d\left(e^{\beta t} X_t\right) = \sigma e^{\beta t} dB_t$$

Proof

Step 1 (Multiply by the integrating factor)
Multiply both sides of the SDE by $\mu(t) = e^{\beta t}$:

$$e^{\beta t} dX_t = -\beta e^{\beta t} X_t \, dt + \sigma e^{\beta t} dB_t$$

But using (Itô's) product rule:

$$d\left(e^{\beta t} X_t\right) = e^{\beta t} dX_t + \beta e^{\beta t} X_t \, dt$$

So we get:

$$d\left(e^{\beta t} X_t\right) = \sigma e^{\beta t} dB_t$$

Step 2 (Integrate both sides)
Now integrate from 0 to t:

$$e^{\beta t} X_t - X_0 = \sigma \int_0^t e^{\beta s} \, dB_s$$

Proof

Step 1 (Multiply by the integrating factor)
Multiply both sides of the SDE by $\mu(t) = e^{\beta t}$:

$$e^{\beta t} dX_t = -\beta e^{\beta t} X_t \, dt + \sigma e^{\beta t} dB_t$$

But using (Itô's) product rule:

$$d\left(e^{\beta t} X_t\right) = e^{\beta t} dX_t + \beta e^{\beta t} X_t \, dt$$

So we get:

$$d\left(e^{\beta t} X_t\right) = \sigma e^{\beta t} dB_t$$

Step 2 (Integrate both sides)
Now integrate from 0 to t:

$$e^{\beta t} X_t - X_0 = \sigma \int_0^t e^{\beta s} \, dB_s$$

Rewriting:

$$X_t = e^{-\beta t} X_0 + \sigma e^{-\beta t} \int_0^t e^{\beta s} \, dB_s$$

Proof (continued)

Step 3 (Distribution of the integral term)
Let: $I_t := \int_0^t e^{\beta s}\, dB_s$. This is an Itô integral of a deterministic function $\Rightarrow$ it's a **Gaussian random variable** with:

Proof (continued)

Step 3 (Distribution of the integral term)
Let: $I_t := \int_0^t e^{\beta s}\, dB_s$. This is an Itô integral of a deterministic function $\Rightarrow$ it's a **Gaussian random variable** with:

- Mean: $\mathbb{E}[I_t] = 0$

Proof (continued)

Step 3 (Distribution of the integral term)
Let: $I_t := \int_0^t e^{\beta s}\, dB_s$. This is an Itô integral of a deterministic function ⇒ it's a **Gaussian random variable** with:

- Mean: $\mathbb{E}[I_t] = 0$
- Variance :

$$\begin{aligned}\mathrm{Var}(I_t) &= \mathbb{E}\left[\left(\int_0^t e^{\beta s}\, dB_s\right)^2\right] = \int_0^t \left(e^{\beta s}\right)^2 ds \quad \text{(using } \textbf{Itô isometry}\text{)} \\ &= \int_0^t e^{2\beta s}\, ds = \left[\frac{1}{2\beta} e^{2\beta s}\right]_0^t = \frac{1}{2\beta}(e^{2\beta t} - 1).\end{aligned}$$

Proof (continued)

Step 3 (Distribution of the integral term)
Let: $I_t := \int_0^t e^{\beta s}\, dB_s$. This is an Itô integral of a deterministic function $\Rightarrow$ it's a **Gaussian random variable** with:

- Mean: $\mathbb{E}[I_t] = 0$
- Variance :

$$\begin{aligned} \mathrm{Var}(I_t) &= \mathbb{E}\left[\left(\int_0^t e^{\beta s}\, dB_s\right)^2\right] = \int_0^t \left(e^{\beta s}\right)^2 ds \quad \text{(using \textbf{Itô isometry})} \\ &= \int_0^t e^{2\beta s}\, ds = \left[\frac{1}{2\beta} e^{2\beta s}\right]_0^t = \frac{1}{2\beta}(e^{2\beta t} - 1). \end{aligned}$$

Therefore:

$$\sigma e^{-\beta t} I_t \sim \mathcal{N}\left(0,\ \sigma^2 e^{-2\beta t} \cdot \frac{1}{2\beta}(e^{2\beta t} - 1)\right) = \mathcal{N}\left(0,\ \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right).$$

Proof (continued)

Step 3 (Distribution of the integral term)
Let: $I_t := \int_0^t e^{\beta s}\, dB_s$. This is an Itô integral of a deterministic function $\Rightarrow$ it's a **Gaussian random variable** with:

- Mean: $\mathbb{E}[I_t] = 0$
- Variance :

$$\mathrm{Var}(I_t) = \mathbb{E}\left[\left(\int_0^t e^{\beta s}\, dB_s\right)^2\right] = \int_0^t \left(e^{\beta s}\right)^2 ds \quad \text{(using \textbf{Itô isometry})}$$

$$= \int_0^t e^{2\beta s}\, ds = \left[\frac{1}{2\beta} e^{2\beta s}\right]_0^t = \frac{1}{2\beta}(e^{2\beta t} - 1).$$

Therefore:

$$\sigma e^{-\beta t} I_t \sim \mathcal{N}\left(0,\ \sigma^2 e^{-2\beta t} \cdot \frac{1}{2\beta}(e^{2\beta t} - 1)\right) = \mathcal{N}\left(0,\ \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right).$$

So the full solution is : $X_t = e^{-\beta t} X_0 + \sigma e^{-\beta t} I_t$, where
$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right)$. **Done!**

(Very) Important remarks

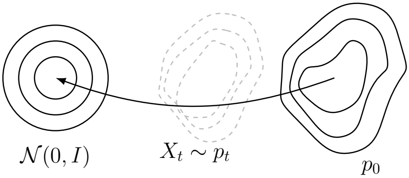


Figure: Representing X_t an OU process (with $\beta = 1$, $\sigma = \sqrt{2}$), and p_t its (time) marginals

- We know that the full solution :

$$X_t = e^{-\beta t} X_0 + \text{Gaussian noise} \tag{2}$$

where Gaussian noise $\sim \mathcal{N}\left(0,\ \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right)$ and that conditionally on X_0:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right) \tag{3}$$

(Very) Important remarks

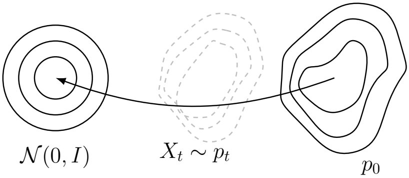


Figure: Representing X_t an OU process (with $\beta = 1$, $\sigma = \sqrt{2}$), and p_t its (time) marginals

- We know that the full solution :

$$X_t = e^{-\beta t} X_0 + \text{Gaussian noise} \tag{2}$$

where Gaussian noise $\sim \mathcal{N}\left(0,\ \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right)$ and that conditionally on X_0:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right) \tag{3}$$

- The marginals $(p_t)_{t \geq 0}$, where p_t the law of X_t in (2) are not Gaussian in general !! (see gray density in the figure above)

(Very) Important remarks

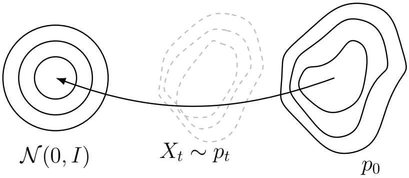


Figure: Representing X_t an OU process (with $\beta = 1$, $\sigma = \sqrt{2}$), and p_t its (time) marginals

- We know that the full solution :

$$X_t = e^{-\beta t} X_0 + \text{Gaussian noise} \tag{2}$$

  where Gaussian noise $\sim \mathcal{N}\left(0,\ \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right)$ and that conditionally on X_0:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right) \tag{3}$$

- The marginals $(p_t)_{t \geq 0}$, where p_t the law of X_t in (2) are not Gaussian in general !! (see gray density in the figure above)
- but the **conditional laws** in (3) are Gaussian

Introducing some initial Condition

When are the marginals p_t Gaussian? Answer: when p_0 is Gaussian.

Introducing some initial Condition

When are the marginals p_t Gaussian? Answer: when p_0 is Gaussian.

Assume $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$.

Then we have $\Rightarrow X_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$.

Proof: Recall $X_t = A + B$ where $A = e^{-\beta t} X_0, \quad B = \sigma e^{-\beta t} \int_0^t e^{\beta s} dW_s$.

- $A \sim \mathcal{N}(0, e^{-2\beta t} \cdot \frac{\sigma^2}{2\beta})$
- $B \sim \mathcal{N}(0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}))$
- $A \perp B \Rightarrow A + B \sim \mathcal{N}(0, \text{sum of variances})$

Introducing some initial Condition

When are the marginals p_t Gaussian? Answer: when p_0 is Gaussian.

Assume $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$.

Then we have $\Rightarrow X_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$.

Proof: Recall $X_t = A + B$ where $A = e^{-\beta t} X_0, \quad B = \sigma e^{-\beta t} \int_0^t e^{\beta s} dW_s$.

- $A \sim \mathcal{N}(0, e^{-2\beta t} \cdot \frac{\sigma^2}{2\beta})$
- $B \sim \mathcal{N}(0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}))$
- $A \perp B \Rightarrow A + B \sim \mathcal{N}(0, \text{sum of variances})$

Above, the law of X_t does not depend on time, because we have started the process at the stationary distribution $\pi(x) = \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$:

$$\text{If: } X_0 \sim \pi(x) \Rightarrow X_t \sim \pi(x) \quad \text{for all } t$$

Introducing some initial Condition

When are the marginals p_t Gaussian? Answer: when p_0 is Gaussian.

Assume $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$.

Then we have $\Rightarrow X_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$.

Proof: Recall $X_t = A + B$ where $A = e^{-\beta t} X_0$, $B = \sigma e^{-\beta t} \int_0^t e^{\beta s} dW_s$.

- $A \sim \mathcal{N}(0, e^{-2\beta t} \cdot \frac{\sigma^2}{2\beta})$
- $B \sim \mathcal{N}(0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}))$
- $A \perp B \Rightarrow A + B \sim \mathcal{N}(0, \text{sum of variances})$

Above, the law of X_t does not depend on time, because we have started the process at the stationary distribution $\pi(x) = \mathcal{N}\left(0, \frac{\sigma^2}{2\beta}\right)$:

$$\text{If: } X_0 \sim \pi(x) \Rightarrow X_t \sim \pi(x) \quad \text{for all } t$$

In general, for a $X_0 \sim \mathcal{N}(0, \sigma_0^2)$, we would have

$$X_t \sim \mathcal{N}\left(0,\ e^{-2\beta t}\sigma_0^2 + \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right).$$

Back to general Langevin diffusion

- We have spent quite a lot of time on Ornstein-Uhlenbeck (OU):

$$dx_t = -\beta x_t \, dt + \sigma \, dB_t$$

Solution:

$$x_t = e^{-\beta t} x_0 + \sigma e^{-\beta t} \int_0^t e^{\beta s} \, dB_s$$

Distribution:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}) \right)$$

Back to general Langevin diffusion

- We have spent quite a lot of time on Ornstein-Uhlenbeck (OU):

$$dx_t = -\beta x_t \, dt + \sigma \, dB_t$$

Solution:

$$x_t = e^{-\beta t} x_0 + \sigma e^{-\beta t} \int_0^t e^{\beta s} \, dB_s$$

Distribution:

$$X_t \mid X_0 \sim \mathcal{N}\left(e^{-\beta t} X_0, \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})\right)$$

- Let's go back to a general Langevin diffusion :

$$\mathrm{d}x_t = -\nabla V(x_t) dt + \sqrt{2} \mathrm{d}B_t, \quad x_t \sim p_t$$

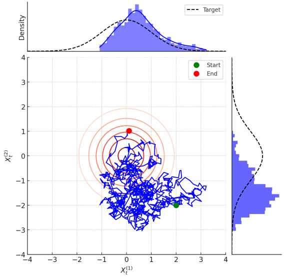
Solution:

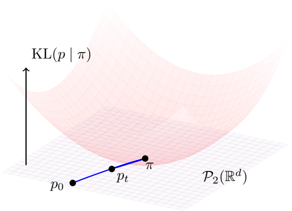
$$x_t = x_0 - \int_0^t \nabla V(x_s) ds + \sqrt{2} B_t,$$

- Remember that OU is a specific case of Langevin, where the target/stationary distribution is: $\pi = \mathcal{N}(0, \frac{\sigma^2}{2\beta})$, where $\pi(x) \propto \exp(-\frac{\beta \|x\|^2}{\sigma^2})$
- **for general Langevin, the stationary distribution is** $\pi \propto \exp(-V)$.

Langevin diffusion (and its discretized versions) is an example of a non-parametric method: we built a process $x_t \in \mathbb{R}^d$, whose distribution p_t converges to π as $t \to \infty$

- The law $(p_t)_{t \geq 0}$ of Langevin diffusion $(x_t)_{t \geq 0}$ is known to follow a gradient flow to minimize $\mathrm{D}(p|\pi) = \mathrm{KL}(p|\pi)$: $\mathrm{d}p_t = -\nabla_{W_2} \mathrm{KL}(p_t|\pi)\mathrm{d}t$ (see [1])



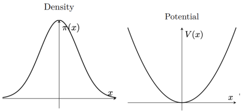


Recall above we plot $x_{t+1} = x_t + \gamma \nabla \log \pi(x_t) + \sqrt{2\gamma}\eta_t$ for $\pi \propto \exp(-\frac{\|x\|^2}{2})$, $x_0 \sim p_0$.

[1] Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the Fokker–Planck equation. SIAM journal on mathematical analysis.

When does Langevin diffusion's law converges (fast) to π?

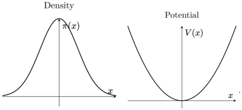
- Consider a standard Gaussian distribution $\pi(x) \propto \exp(-\frac{\|x\|^2}{2})$, i.e. $\pi \propto \exp(-V)$ with V 1-strongly convex, i.e. π is (1-)strongly log-concave.



Then $\mathrm{KL}(p_t|\pi) = \exp(-2t)\,\mathrm{KL}(p_0|\pi)$.

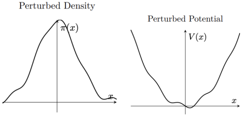
When does Langevin diffusion's law converges (fast) to π?

- Consider a standard Gaussian distribution $\pi(x) \propto \exp(-\frac{\|x\|^2}{2})$, i.e. $\pi \propto \exp(-V)$ with V 1-strongly convex, i.e. π is (1-)strongly log-concave.



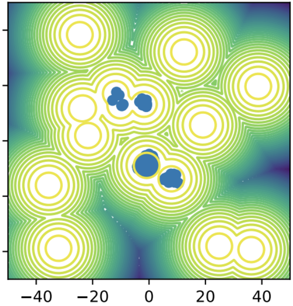
Then $\mathrm{KL}(p_t|\pi) = \exp(-2t)\,\mathrm{KL}(p_0|\pi)$.

- If π is a perturbation of a strongly-log-concave distribution, then the rate degrades with the size of the perturbation.



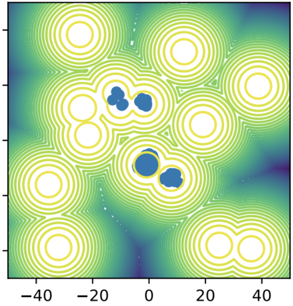
(see Holley–Stroock theorem and log-Sobolev inequalities, (Bakry et al., 2014)).

Langevin in the multimodal case



Mixture of equally weighted 16 Gaussians with unit variance and uniformly chosen centers in $[-40, 40]^2$, a standard sampling benchmark. ULA was initialized with $\mathcal{N}(0, I_2)$, step-size $h = 0.01$. ULA was run with 5.10^4 steps (one minute run).

Langevin in the multimodal case



Mixture of equally weighted 16 Gaussians with unit variance and uniformly chosen centers in $[-40, 40]^2$, a standard sampling benchmark. ULA was initialized with $\mathcal{N}(0, I_2)$, step-size $h = 0.01$. ULA was run with 5.10^4 steps (one minute run).

The theoretical convergence is so slow, that in practice Langevin gets stuck for infinite time the modes close to its initialization !

Outline

Bayesian learning

Langevin

Bayesian deep learning

Recall Bayesian inference

Given labelled data $(w_i, y_i)_{i=1}^p$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i,\, y_i)_{i=1}^p} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

I.e., $\pi(x) = \frac{\exp(-V(x))}{Z}$, $V(x) = -\log p(\mathcal{D}|x) - \log p_0(x)$ with Z intractable.

Recall Bayesian inference

Given labelled data $(w_i, y_i)_{i=1}^{p}$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

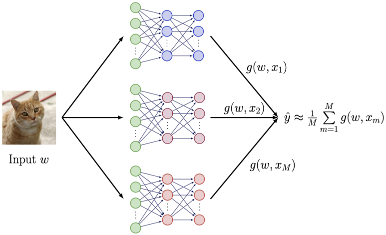
$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i, y_i)_{i=1}^{p}} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

I.e., $\pi(x) = \frac{\exp(-V(x))}{Z}$, $V(x) = -\log p(\mathcal{D}|x) - \log p_0(x)$ with Z intractable.

Ensemble prediction for an input w:

$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w, x) d\pi(x)}_{\text{"Bayesian model averaging"}}$$

Predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\pi(x)$.



Recall Bayesian inference

Given labelled data $(w_i, y_i)_{i=1}^{p}$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

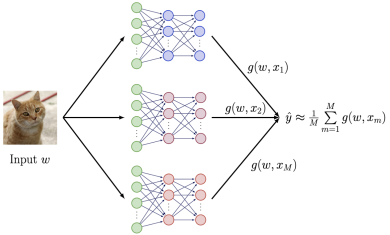
$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i, y_i)_{i=1}^{p}} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

I.e., $\pi(x) = \frac{\exp(-V(x))}{Z}$, $V(x) = -\log p(\mathcal{D}|x) - \log p_0(x)$ with Z intractable.

Ensemble prediction for an input w:

$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w,x) d\pi(x)}_{\text{"Bayesian model averaging"}}$$

Predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\pi(x)$.



recall that a frequentist NN would predict $\hat{y} = g(w, x^*)$ where $x^* = \arg\max_{x \in \mathbb{R}^d} \log p(\mathcal{D}|x)$

Langevin for (Bayesian) deep NN?

Given labelled data $\mathcal{D} = (w_i, y_i)_{i=1}^p$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i, y_i)_{i=1}^p} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

Langevin for (Bayesian) deep NN?

Given labelled data $\mathcal{D} = (w_i, y_i)_{i=1}^p$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i, y_i)_{i=1}^p} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

- Recall that we know that the convergence speed of Langevin diffusion depends on how much "V is convex" and if it has few local minimas

Langevin for (Bayesian) deep NN?

Given labelled data $\mathcal{D} = (w_i, y_i)_{i=1}^p$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i, y_i)_{i=1}^p} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

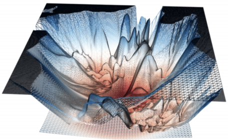
- Recall that we know that the convergence speed of Langevin diffusion depends on how much "V is convex" and if it has few local minimas
- is $x \mapsto V(x)$ convex for $g(., x)$ a neural network parametrized by x?

Langevin for (Bayesian) deep NN?

Given labelled data $\mathcal{D} = (w_i, y_i)_{i=1}^p$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i,\, y_i)_{i=1}^p} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}} .$$

- Recall that we know that the convergence speed of Langevin diffusion depends on how much "V is convex" and if it has few local minimas
- is $x \mapsto V(x)$ convex for $g(., x)$ a neural network parametrized by x?



A highly nonconvex loss surface, as is common in deep neural nets. From `https://www.telesens.co/2019/01/16/neural-network-loss-visualization`.

Different strategies in practice/in the literature

Close to what we've seen previously:

- Stochastic Langevin dynamics: approximate $\nabla V(x) = \nabla \left(\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2 + \frac{\|x\|^2}{2} \right)$ by a batch of data samples $(w_i, y_i)_{i=1}^{m}$ with $m << p$
- Variational Inference

$$\text{find } q_\theta = \arg\min_{p \in P_\theta} \mathrm{KL}(p|\pi)$$

where P_θ is a family of parametric distributions (upcoming in few slides).

Different strategies in practice/in the literature

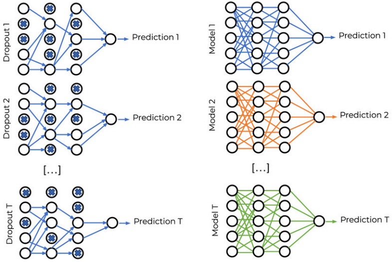
More heuristic:

- Monte Carlo Dropout

  Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning.

- Deep ensembles

  Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems.



(a) MC Dropout (b) Ensemble Method

Variational Inference for BNN - Bayes by Backprop example

Variational Inference

$$\text{find } q_\theta = \arg\min_{p \in P_\theta} \mathrm{KL}(p|\pi)$$

where P_θ is a family of parametric distributions.

Variational Inference for BNN - Bayes by Backprop example

Variational Inference

$$\text{find } q_\theta = \underset{p \in P_\theta}{\arg\min}\, \mathrm{KL}(p|\pi)$$

where P_θ is a family of parametric distributions.

A typical neural network of depth L (with non-linearity $h(\cdot)$) for input w and parameter x writes:

$$g(w,x) = A^L h\left(A^{L-1} h\left(\dots h\left(A^1 w + b^1\right)\right) + b^{L-1}\right) + b^L,$$

$$h^l = h(A^l h^{l-1} + b^l), \quad h^1 = h(A^1 w + b^1).$$

Neural network parameters: $x = \{A^l, b^l\}_{l=1}^L$.

We will describe the approach of "**Bayes by Backprop**"[1].

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In International conference on machine learning.

Step 1: Construct the $q_\theta(x) \approx p(x \mid \mathcal{D}) = \pi(x)$ Distribution

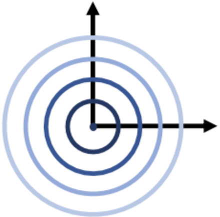
Example: Mean-field (="factorized") Gaussian distribution:

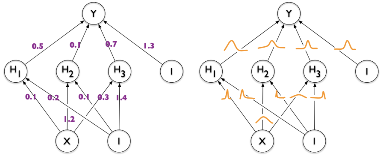
$$q_\theta = \prod_{l=1}^{L} q(A^l)\, q(b^l)$$

$$q(A_l) = \prod_{ij} q(A^l_{ij}), \quad q(A^l_{ij}) = \mathcal{N}(A^l_{ij}; M^l_{ij}, V^l_{ij})$$

$$q(b^l) = \prod_{i} q(b^l_{i}), \quad q(b^l_{i}) = \mathcal{N}(b^l_{i}; m^l_{i}, v^l_{i})$$

Variational parameters: $\theta = \left\{M^l_{ij}, V^l_{ij}, m^l_i, v^l_i\right\}_{l=1}^{L}$





In dimension two, a simple example of q_θ is a factorized Gaussian:

$$q_\theta(A^1_{11}, A^1_{12}) = \mathcal{N}(A^1_{11}; 0, 1) \cdot \mathcal{N}(A^1_{12}; 0, 1),$$

where q_θ is the product of two independent standard normal distributions over the parameters A^1_{11} and A^1_{12}.

Note that the "factor" assumption in mean-field decorrelates variables.

Step 2: Fit the q_θ Distribution

Variational inference: $\theta^* = \arg\max L(\theta)$ where L is the ELBO

$$L(\theta) = \mathbb{E}_{q_\theta}[\log p(D \mid x)] - \mathrm{KL}[q_\theta \parallel p_0(x)]$$

First scalable technique: Stochastic optimization

- i.i.d. assumption: $\log p(D \mid x) = \sum_{i=1}^{N} \log p(y_i \mid w_i, x)$
- Mini-batch training: $\{(w_m, y_m)\}_{m=1}^{M} \sim D^M$

$$L(\theta) \approx \frac{N}{M} \sum_{i=1}^{M} \mathbb{E}_{q_\theta}[\log p(y_i \mid w_i, x)] - \mathrm{KL}[q_\theta \parallel p_0(x)]$$

Reweighting to ensure calibrated posterior concentration.

Step 2: Fit the q_θ Distribution

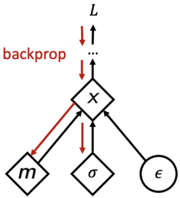
Variational inference: $\theta^* = \arg\max L(\theta)$ where L is the ELBO

$$L(\theta) = \mathbb{E}_{q_\theta}[\log p(D \mid x)] - \mathrm{KL}[q_\theta \,\|\, p_0(x)]$$

2nd Scalable Technique: Monte Carlo Sampling

- $\mathbb{E}_{q_\theta}[\log p(y \mid w, x)]$ is intractable even with Gaussian q_θ
- Solution: Monte Carlo estimate:

$$\mathbb{E}_{q_\theta}[\log p(y \mid w, x)] \approx \frac{1}{K}\sum_{k=1}^{K} \log p(y \mid w, x_k), \quad x_k \sim q_\theta$$



- Reparameterization trick to sample from mean-field Gaussians:

$$x_k = m_\theta + \sigma_\theta \odot \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, I)$$

- Therefore:

$$\mathbb{E}_{q_\theta}[\log p(y \mid w, x)] \approx \frac{1}{K}\sum_{k=1}^{K} \log p(y \mid w, x_k), \; x_k = m_\theta + \sigma_\theta \epsilon_k$$

Combining both steps and final prediction

Full ELBO approximation:

$$L(\theta) \approx \frac{N}{M} \sum_{m=1}^{M} \frac{1}{K} \sum_{k=1}^{K} \log p(y_m \mid w_m, x_k) - \mathrm{KL}[q_\theta \,\|\, p(x)], \quad x_k \sim q_\theta$$

analytic between two Gaussians (if not, can also be estimated with Monte Carlo)

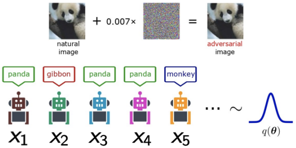
In regression: $p(y \mid w, x) = \mathcal{N}(f_x(w), \sigma^2)$,

In classification: $p(y \mid w, x) = \mathrm{Categorical}(\mathrm{logit} = f_x(w))$

Step 3: Compute Prediction with Monte Carlo Approximations

$$p(y^* \mid w^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid w^*, x_k), \quad x_k \sim q_\theta$$

Mean-field Gaussian case: $x_k = m_\theta + \sigma_\theta \odot \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, I)$



References I

Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of Markov diffusion operators*, volume 103. Springer.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.