Convex Optimization 10-725, Lecture 30: Optimization for Sampling.

Yuanzhi Li

Assistant Professor, Carnegie Mellon University
Visiting Researcher, Microsoft Research

Today

So Far

- We have been focusing on minimizing function $f(x)$ under different conditions.

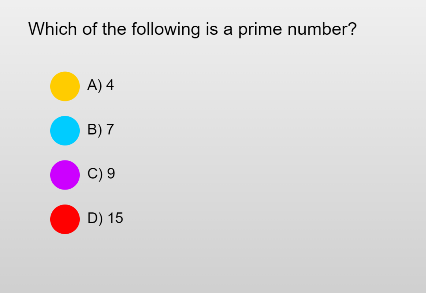
So Far

- We have been focusing on minimizing function $f(x)$ under different conditions.
- This is mainly useful for the discriminative machine learning approach.

So Far

- We have been focusing on minimizing function $f(x)$ under different conditions.
- This is mainly useful for the discriminative machine learning approach.
- Discriminative: We are given a bunch of observations, and we want to predict their labels.

So Far

- We have been focusing on minimizing function $f(x)$ under different conditions.
- This is mainly useful for the discriminative machine learning approach.
- Discriminative: We are given a bunch of observations, and we want to predict their labels.
- 

So Far

- Discriminative: We are given a bunch of observations, and we want to predict their labels.

So Far

- Discriminative: We are given a bunch of observations, and we want to predict their labels.
- The empirical risk minimization objective:

$$\min_{w} f(w) = \sum_{i \in [N]} \ell(h(w, x^{(i)}), y^{(i)}) + R(w)$$

So Far

- Discriminative: We are given a bunch of observations, and we want to predict their labels.
- The empirical risk minimization objective:

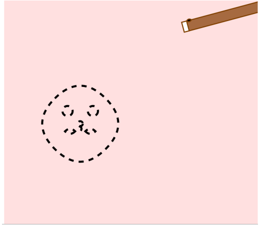
$$\min_w f(w) = \sum_{i \in [N]} \ell(h(w, x^{(i)}), y^{(i)}) + R(w)$$

- However, discriminative is not the only way of learning. The other way is "generative".

So Far

- Discriminative: We are given a bunch of observations, and we want to predict their labels.
- The empirical risk minimization objective:

$$\min_w f(w) = \sum_{i \in [N]} \ell(h(w, x^{(i)}), y^{(i)}) + R(w)$$

- However, discriminative is not the only way of learning. The other way is "generative".
- 

So Far

- Generative model: Given a bunch of observations, we want to generate new observations that are sampled from the same distribution.

So Far

- Generative model: Given a bunch of observations, we want to generate new observations that are sampled from the same distribution.
- Mathematically, the question is now formulated as a distribution learning problem.

So Far

- Generative model: Given a bunch of observations, we want to generate new observations that are sampled from the same distribution.
- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:

So Far

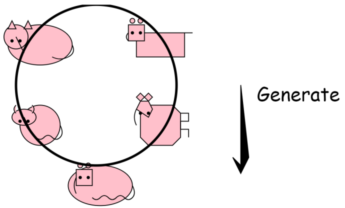
- Generative model: Given a bunch of observations, we want to generate new observations that are sampled from the same distribution.
- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:
- Generative approach: Can we generate new samples $x \sim \mathcal{D}$?

So Far

- Generative model: Given a bunch of observations, we want to generate new observations that are sampled from the same distribution.
- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:
- Generative approach: Can we generate new samples $x \sim \mathcal{D}$?
- For example, can the machines learn to generate images given a bunch of images?

So Far

- Generative model: Given a bunch of observations, we want to generate new observations that are sampled from the same distribution.
- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:
- Generative approach: Can we generate new samples $x \sim \mathcal{D}$?
- For example, can the machines learn to generate images given a bunch of images?



-

Sampling

- Mathematically, the question is now formulated as a distribution learning problem.

Sampling

- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:

Sampling

- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:
- Can we generate new samples $x \sim \mathcal{D}$?

Sampling

- Mathematically, the question is now formulated as a distribution learning problem.
- Given observations $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$, i.i.d. sampled from some unknown distribution $\mathcal{D}$:
- Can we generate new samples $x \sim \mathcal{D}$?
- Key question for this lecture: How do we optimize to learn a distribution?

Sampling

- Key question for this lecture: How do we optimize to learn a distribution?

Sampling

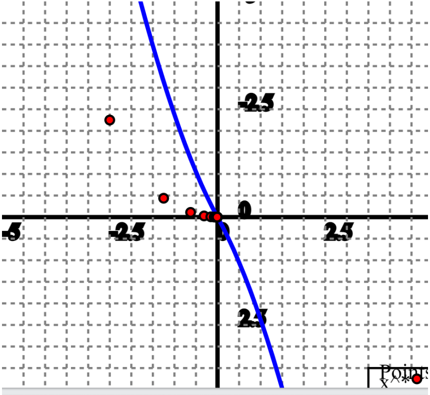
- Key question for this lecture: How do we optimize to learn a distribution?
- Recall the optimization algorithms we have learned so far:

Sampling

- Key question for this lecture: How do we optimize to learn a distribution?
- Recall the optimization algorithms we have learned so far:
- Given an arbitrary point x_0, our goal is to create a sequence of points $x_1, x_2, \cdots, x_t$ that gets closer and closer to x^*.

Sampling

- Key question for this lecture: How do we optimize to learn a distribution?
- Recall the optimization algorithms we have learned so far:
- Given an arbitrary point x_0, our goal is to create a sequence of points $x_1, x_2, \cdots, x_t$ that gets closer and closer to x^*.



-

Sampling

- Key question for this lecture: How do we optimize to learn a distribution?

Sampling

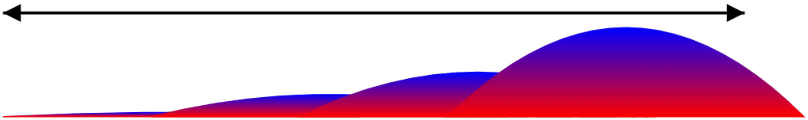
- Key question for this lecture: How do we optimize to learn a distribution?
- Our goal is to get closer to the target distribution.

Sampling

- Key question for this lecture: How do we optimize to learn a distribution?
- Our goal is to get closer to the target distribution.
- Apply the same way of thinking, can we create a sequence of distributions q_t that gets closer and closer to q^*, where q_0 is some simple distribution (such as standard Gaussian)

Sampling

- Key question for this lecture: How do we optimize to learn a distribution?
- Our goal is to get closer to the target distribution.
- Apply the same way of thinking, can we create a sequence of distributions q_t that gets closer and closer to q^*, where q_0 is some simple distribution (such as standard Gaussian)



Sampling

- Can we create a sequence of distributions q_t that gets closer and closer to the target distribution q^*, where q_0 is some simple distribution (such as standard Gaussian)?

Sampling

- Can we create a sequence of distributions q_t that gets closer and closer to the target distribution q^*, where q_0 is some simple distribution (such as standard Gaussian)?
- We will talk about the so-called diffusion approach.

Diffusion optimization

- Key question: How do we create a sequence of distribution that gets closer and closer to the target distribution q^*, where q_0 is standard Gaussian?

Diffusion optimization

- Key question: How do we create a sequence of distribution that gets closer and closer to the target distribution q^*, where q_0 is standard Gaussian?
- Intuitively:

Diffusion optimization

- Key question: How do we create a sequence of distribution that gets closer and closer to the target distribution q^*, where q_0 is standard Gaussian?
- Intuitively:
- We first create a sequence that starts from q^*, and make it get closer and closer to standard Gaussian as we increase t.

Diffusion optimization

- Key question: How do we create a sequence of distribution that gets closer and closer to the target distribution q^*, where q_0 is standard Gaussian?
- Intuitively:
- We first create a sequence that starts from q^*, and make it get closer and closer to standard Gaussian as we increase t.
- We then reverse the process to go from Gaussian to q^*.

Diffusion optimization

- Going from a target distribution q^* to Gaussian:

Diffusion optimization

- Going from a target distribution q^* to Gaussian:
- Simple, we can use the so called Brownian motion.

Diffusion optimization

- Going from a target distribution q^* to Gaussian:
- Simple, we can use the so called Brownian motion.
- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

Diffusion optimization

- Going from a target distribution q^* to Gaussian:
- Simple, we can use the so called Brownian motion.
- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

- Theorem (classical result): As $t \to \infty$, the distribution of $X_t \to \mathcal{N}(0, I)$.

Diffusion optimization

- Going from a target distribution q^* to Gaussian:
- Simple, we can use the so called Brownian motion.
- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

- Theorem (classical result): As $t \to \infty$, the distribution of $X_t \to \mathcal{N}(0, I)$.
- In fact, we even know the explicit form: ($\gamma \sim \mathcal{N}(0, I)$ is the standard Gaussian)

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

Diffusion optimization

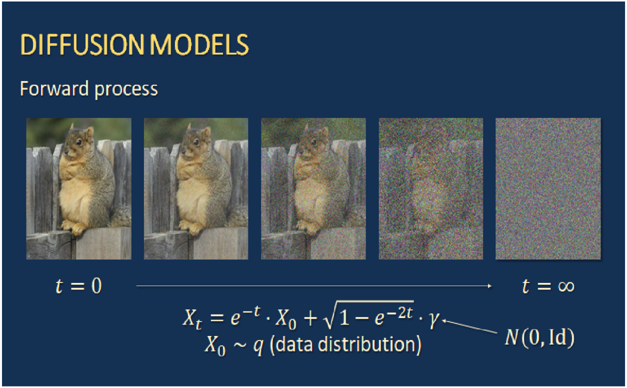
- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

Diffusion optimization

- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

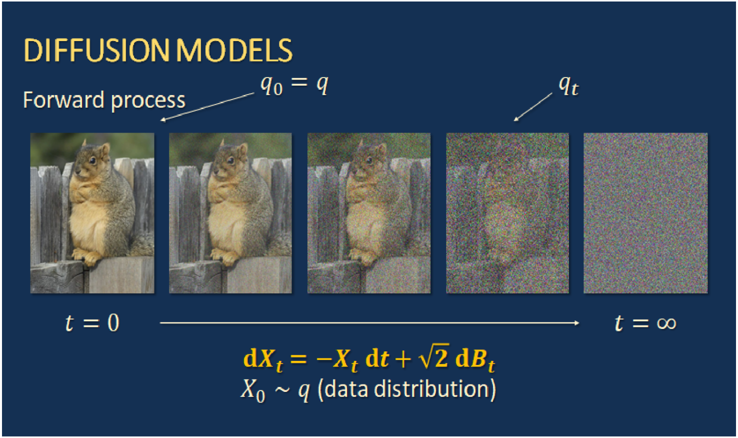
$$dX_t = -X_t dt + \sqrt{2} dB_t$$

- 

Diffusion optimization

- How do we go backward? – Meaning that how do we go from $X_\infty \to X_0$?

Diffusion optimization

- How do we go backward? – Meaning that how do we go from $X_\infty \to X_0$?
- 

Diffusion optimization

- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

Diffusion optimization

- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

- By Fokker–Planck equation:

More generally, if

$$d\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t, t)\, dt + \boldsymbol{\sigma}(\mathbf{X}_t, t)\, d\mathbf{W}_t,$$

where $\mathbf{X}_t$ and $\boldsymbol{\mu}(\mathbf{X}_t, t)$ are N-dimensional random vectors, $\boldsymbol{\sigma}(\mathbf{X}_t, t)$ is an $N \times M$ matrix and $\mathbf{W}_t$ is an M-dimensional standard Wiener process, the probability density $p(\mathbf{x}, t)$ for $\mathbf{X}_t$ satisfies the Fokker–Planck equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} [\mu_i(\mathbf{x}, t) p(\mathbf{x}, t)] + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial^2}{\partial x_i \, \partial x_j} [D_{ij}(\mathbf{x}, t) p(\mathbf{x}, t)],$$

with drift vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and diffusion tensor $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$, i.e.

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2} \sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t) \sigma_{jk}(\mathbf{x}, t).$$

Let $q_t(X)$ be the density function of X_t, then we have: for every X,

$$\frac{\partial q_t(X)}{\partial t} := \langle \nabla, X q_t(X) \rangle + \langle \nabla^2 q_t(X), E \rangle$$

Here E is the all one matrix. ∇ is taken w.r.t. X

Diffusion optimization

- Starting from $X_0 \sim q^*$, we can apply: (B_t is the brownian motion, you can think dB_t as adding independent, standard Gaussian noise.)

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

- By Fokker–Planck equation:

More generally, if

$$d\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t, t)\,dt + \boldsymbol{\sigma}(\mathbf{X}_t, t)\,d\mathbf{W}_t,$$

where $\mathbf{X}_t$ and $\boldsymbol{\mu}(\mathbf{X}_t, t)$ are N-dimensional random vectors, $\boldsymbol{\sigma}(\mathbf{X}_t, t)$ is an $N \times M$ matrix and $\mathbf{W}_t$ is an M-dimensional standard Wiener process, the probability density $p(\mathbf{x}, t)$ for $\mathbf{X}_t$ satisfies the Fokker–Planck equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[\mu_i(\mathbf{x}, t) p(\mathbf{x}, t)\right] + \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2}{\partial x_i\,\partial x_j}\left[D_{ij}(\mathbf{x}, t) p(\mathbf{x}, t)\right],$$

with drift vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and diffusion tensor $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$, i.e.

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2}\sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t)\sigma_{jk}(\mathbf{x}, t).$$

Let $q_t(X)$ be the density function of X_t, then we have: for every X,

$$\frac{\partial q_t(X)}{\partial t} := \langle \nabla, X q_t(X) \rangle + \langle \nabla^2 q_t(X), E \rangle$$

Here E is the all one matrix. ∇ is taken w.r.t. X

- In the forward process, $q_0 = q^*$ is the target distribution, $q_T \to$ standard Gaussian as $T \to \infty$.

Diffusion optimization

- In the forward process, $q_0 = q^*$ is the target distribution, $q_T \to$ standard Gaussian as $T \to \infty$.

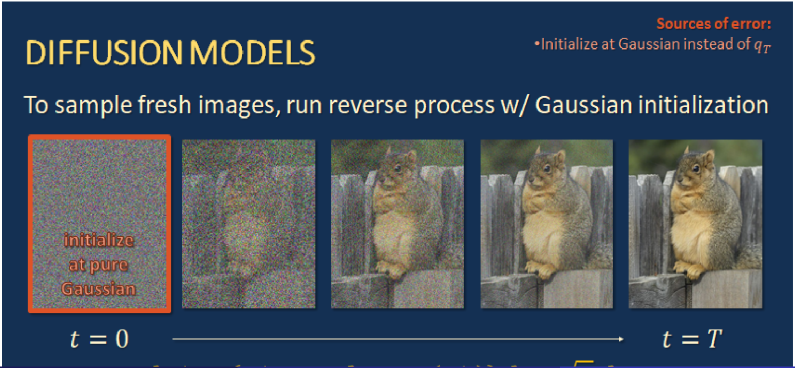
Diffusion optimization

- In the forward process, $q_0 = q^*$ is the target distribution, $q_T \to$ standard Gaussian as $T \to \infty$.
- Now, suppose we want to have a backward process, where $p_t(X) = q_{T-t}(X)$ for sufficiently large T.

Diffusion optimization

- In the forward process, $q_0 = q^*$ is the target distribution, $q_T \to$ standard Gaussian as $T \to \infty$.
- Now, suppose we want to have a backward process, where $p_t(X) = q_{T-t}(X)$ for sufficiently large T.
- In this process, p_0 will be (close to) standard Gaussian, and $p_T = q^*$ as the target distribution.

Diffusion optimization

- In the forward process, $q_0 = q^*$ is the target distribution, $q_T \to$ standard Gaussian as $T \to \infty$.
- Now, suppose we want to have a backward process, where $p_t(X) = q_{T-t}(X)$ for sufficiently large T.
- In this process, p_0 will be (close to) standard Gaussian, and $p_T = q^*$ as the target distribution.
- 

Backward Process

- Main observation: Since the forward process q_t satisfies

$$\frac{\partial q_t(X)}{\partial t} := \langle \nabla, X q_t(X) \rangle + \langle \nabla^2 q_t(X), E \rangle$$

Backward Process

- Main observation: Since the forward process q_t satisfies

$$\frac{\partial q_t(X)}{\partial t} := \langle \nabla, X q_t(X) \rangle + \langle \nabla^2 q_t(X), E \rangle$$

- The backward process $p_t(X) = p_{T-t}(X)$ satisfies:

$$\frac{\partial p_t(X)}{\partial t} = \frac{\partial q_{T-t}(X)}{\partial t} := -\langle \nabla, X q_{T-t}(X) \rangle - \langle \nabla^2 q_{T-t}(X), E \rangle$$

Backward Process

- The backward process $p_t(X) = q_{T-t}(X)$ satisfies:

$$\frac{\partial p_t(X)}{\partial t} := -\langle \nabla, X p_t(X) \rangle - \langle \nabla^2 p_t(X), E \rangle$$

Backward Process

- The backward process $p_t(X) = q_{T-t}(X)$ satisfies:

$$\frac{\partial p_t(X)}{\partial t} := -\langle \nabla, X p_t(X) \rangle - \langle \nabla^2 p_t(X), E \rangle$$

- Look at Fokker–Planck equation again

More generally, if

$$d\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t, t)\, dt + \boldsymbol{\sigma}(\mathbf{X}_t, t)\, d\mathbf{W}_t,$$

where $\mathbf{X}_t$ and $\boldsymbol{\mu}(\mathbf{X}_t, t)$ are N-dimensional random vectors, $\boldsymbol{\sigma}(\mathbf{X}_t, t)$ is an $N \times M$ matrix and $\mathbf{W}_t$ is an M-dimensional standard Wiener process, the probability density $p(\mathbf{x}, t)$ for $\mathbf{X}_t$ satisfies the Fokker–Planck equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} [\mu_i(\mathbf{x}, t) p(\mathbf{x}, t)] + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial^2}{\partial x_i\, \partial x_j} [D_{ij}(\mathbf{x}, t) p(\mathbf{x}, t)],$$

with drift vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and diffusion tensor $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$, i.e.

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2} \sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t) \sigma_{jk}(\mathbf{x}, t).$$

Backward Process

- The backward process $p_t(X) = q_{T-t}(X)$ satisfies:

$$\frac{\partial p_t(X)}{\partial t} := -\langle \nabla, X p_t(X) \rangle - \langle \nabla^2 p_t(X), E \rangle$$

- Look at Fokker–Planck equation again

More generally, if

$$d\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t, t)\, dt + \boldsymbol{\sigma}(\mathbf{X}_t, t)\, d\mathbf{W}_t,$$

where $\mathbf{X}_t$ and $\boldsymbol{\mu}(\mathbf{X}_t, t)$ are N-dimensional random vectors, $\boldsymbol{\sigma}(\mathbf{X}_t, t)$ is an $N \times M$ matrix and $\mathbf{W}_t$ is an M-dimensional standard Wiener process, the probability density $p(\mathbf{x}, t)$ for $\mathbf{X}_t$ satisfies the Fokker–Planck equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left[\mu_i(\mathbf{x}, t) p(\mathbf{x}, t)\right] + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial^2}{\partial x_i\, \partial x_j} \left[D_{ij}(\mathbf{x}, t) p(\mathbf{x}, t)\right],$$

with drift vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and diffusion tensor $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$, i.e.

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2} \sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t) \sigma_{jk}(\mathbf{x}, t).$$

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$

$$dY_t = Y_t + 2\nabla \ln p_t(Y_t) dt + \sqrt{2} dB_t$$

Backward Process

- The backward process $p_t(X) = q_{T-t}(X)$ satisfies:

$$\frac{\partial p_t(X)}{\partial t} := -\langle \nabla, X p_t(X) \rangle - \langle \nabla^2 p_t(X), E \rangle$$

- Look at Fokker–Planck equation again

More generally, if

$$d\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t, t)\, dt + \boldsymbol{\sigma}(\mathbf{X}_t, t)\, d\mathbf{W}_t,$$

where $\mathbf{X}_t$ and $\boldsymbol{\mu}(\mathbf{X}_t, t)$ are N-dimensional random vectors, $\boldsymbol{\sigma}(\mathbf{X}_t, t)$ is an $N \times M$ matrix and $\mathbf{W}_t$ is an M-dimensional standard Wiener process, the probability density $p(\mathbf{x}, t)$ for $\mathbf{X}_t$ satisfies the Fokker–Planck equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left[\mu_i(\mathbf{x}, t) p(\mathbf{x}, t)\right] + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial^2}{\partial x_i\, \partial x_j} \left[D_{ij}(\mathbf{x}, t) p(\mathbf{x}, t)\right],$$

with drift vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and diffusion tensor $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$, i.e.

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2} \sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t) \sigma_{jk}(\mathbf{x}, t).$$

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$

$$dY_t = Y_t + 2\nabla \ln p_t(Y_t) dt + \sqrt{2} dB_t$$

- This is due to

$$2\nabla \ln p_t(Y) = \frac{2\nabla p_t(Y)}{p_t(Y)}$$

Backward Process

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$
$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$

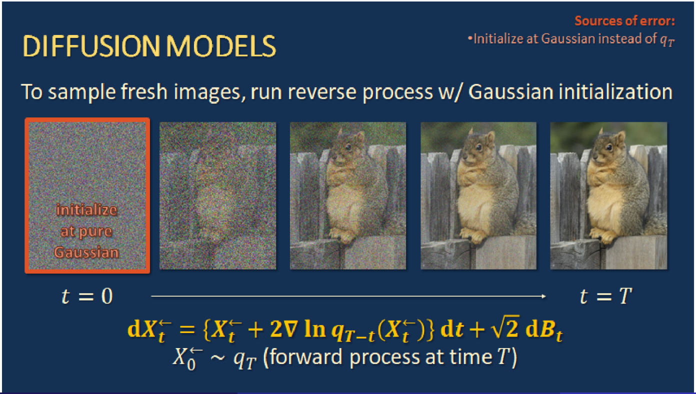
Backward Process

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$
$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$
- This process satisfies that $p_0 \approx$ standard Gaussian, and $p_T = q^*$ is the target distribution.

Backward Process

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$

$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$

- This process satisfies that $p_0 \approx$ standard Gaussian, and $p_T = q^*$ is the target distribution.
- 

Backward Process

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$

$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$

Backward Process

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$
$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$
- This process satisfies that $p_0 \approx$ standard Gaussian, and $p_T = q^*$ is the target distribution.

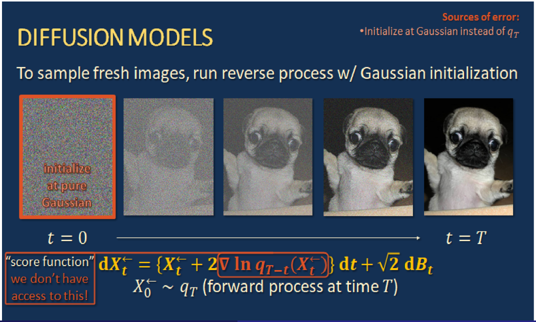
Backward Process

- We know that the backward process is a stochastic process with $(Y_t \sim p_t)$
$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$
- This process satisfies that $p_0 \approx$ standard Gaussian, and $p_T = q^*$ is the target distribution.
- How do we execute this optimization process? – We don't know $\nabla \ln p_t(Y_t)$.

Backward Process

- We know that the backward process is a stochastic process with ($Y_t \sim p_t$)

$$dY_t = Y_t + 2\nabla \ln p_t(Y_t)dt + \sqrt{2}dB_t$$

- This process satisfies that $p_0 \approx$ standard Gaussian, and $p_T = q^*$ is the target distribution.
- How do we execute this optimization process? – We don't know $\nabla \ln p_t(Y_t)$.
- 

Backward Process

- Now we need to estimate $\nabla \ln p_t(Y_t) = \nabla \ln q_{T-t}(Y_t)$.

Backward Process

- Now we need to estimate $\nabla \ln p_t(Y_t) = \nabla \ln q_{T-t}(Y_t)$.
- First idea: Use supervised learning, train a neural network s to minimize:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

Backward Process

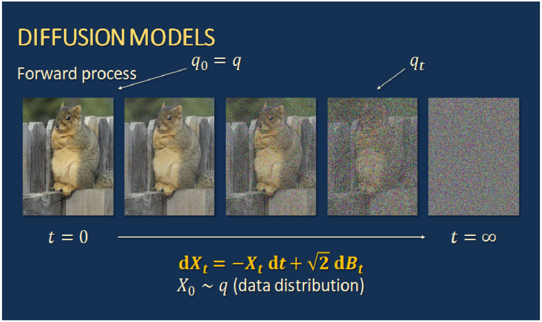
- Now we need to estimate $\nabla \ln p_t(Y_t) = \nabla \ln q_{T-t}(Y_t)$.
- First idea: Use supervised learning, train a neural network s to minimize:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

- However, we only have samples $X \sim q^* = q_0$, how do we compute $\nabla \ln q_t(X_t)$? – This is called score estimation.

- Now we need to estimate $\nabla \ln p_t(Y_t) = \nabla \ln q_{T-t}(Y_t)$.
- First idea: Use supervised learning, train a neural network s to minimize:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

- However, we only have samples $X \sim q^* = q_0$, how do we compute $\nabla \ln q_t(X_t)$? – This is called score estimation.
- 

Backward Process

- We want to minimize:

$$\min_{s} \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

Backward Process

- We want to minimize:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

- Here, $X_0 \sim q_0 = q^*$,

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

Backward Process

- We want to minimize:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

- Here, $X_0 \sim q_0 = q^*$,

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

- Now we do some calculation:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

$$= \min_s \mathbb{E}_{X_t} \left[\| s(X_t) \|^2 - 2 \cdot \langle s(X_t), \nabla \ln q_t(X_t) \rangle + \| \nabla \ln q_t(X_t) \|^2 \right]$$

Backward Process

- We want to minimize:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

- Here, $X_0 \sim q_0 = q^*$,

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

- Now we do some calculation:

$$\min_s \mathbb{E}_{X_t} \| s(X_t) - \nabla \ln q_t(X_t) \|_2^2$$

$$= \min_s \mathbb{E}_{X_t} \left[\| s(X_t) \|^2 - 2 \cdot \langle s(X_t), \nabla \ln q_t(X_t) \rangle + \| \nabla \ln q_t(X_t) \|^2 \right]$$

- We consider the middle term:

$$\mathbb{E}_{X_t} [\langle s(X_t), \nabla \ln q_t(X_t) \rangle] = \int \langle s, \nabla \ln q_t \rangle \, dq_t$$

Backward Process

- We consider the middle term:

$$\mathbb{E}_{X_t}[\langle s(X_t), \nabla \ln q_t(X_t)\rangle] = \int \langle s, \nabla \ln q_t \rangle \, dq_t$$

Backward Process

- We consider the middle term:

$$\mathbb{E}_{X_t}[\langle s(X_t), \nabla \ln q_t(X_t)\rangle] = \int \langle s, \nabla \ln q_t\rangle \, dq_t$$

- Now, we know that

$$\int \langle s, \nabla \ln q_t\rangle \, dq_t = \int \langle s, \nabla \ln q_t\rangle q_t dt = \int \langle s, \nabla q_t\rangle$$

Backward Process

- We consider the middle term:

$$\mathbb{E}_{X_t}[\langle s(X_t), \nabla \ln q_t(X_t)\rangle] = \int \langle s, \nabla \ln q_t \rangle \, dq_t$$

- Now, we know that

$$\int \langle s, \nabla \ln q_t \rangle \, dq_t = \int \langle s, \nabla \ln q_t \rangle q_t dt = \int \langle s, \nabla q_t \rangle$$

- Using integration by part ($\int udv = uv - \int vdu$):

$$\int \langle s, \nabla q_t \rangle = - \int (\nabla \cdot \mathbf{s}) \, dq_t$$

Backward Process

- Using integration by part ($\int udv = uv - \int vdu$):

$$\int \langle s, \nabla q_t \rangle = -\int (\nabla \cdot \mathbf{s})\, dq_t$$

Backward Process

- Using integration by part ($\int udv = uv - \int vdu$):

$$\int \langle s, \nabla q_t \rangle = -\int (\nabla \cdot \mathbf{s})\, dq_t$$

- Now, using the formula of X_t:

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

Backward Process

- Using integration by part ($\int udv = uv - \int vdu$):

$$\int \langle s, \nabla q_t \rangle = - \int (\nabla \cdot \mathbf{s})\, dq_t$$

- Now, using the formula of X_t:

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

- We know that

$$- \int (\nabla \cdot \mathbf{s})\, dq_t = - \iint (\nabla \cdot \mathbf{s})(e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma)\, dq_0(X_0)\, d\gamma$$

Backward Process

- Using integration by part ($\int udv = uv - \int vdu$):

$$\int \langle s, \nabla q_t \rangle = - \int (\nabla \cdot \mathbf{s})\, dq_t$$

- Now, using the formula of X_t:

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

- We know that

$$- \int (\nabla \cdot \mathbf{s})\, dq_t = - \iint (\nabla \cdot \mathbf{s})(e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma)\, dq_0(X_0)\, d\gamma$$

- Remark: $(\nabla \cdot \mathbf{s})(e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma)$ means we compute the value of $\nabla \cdot \mathbf{s}$ at the point $X_t = (e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma)$.

Backward Process

- We know that

$$-\int (\nabla \cdot \mathbf{s})\, dq_t = -\iint (\nabla \cdot \mathbf{s})(e^{-t}X_0 + \sqrt{1-e^{-2t}}\gamma)\, dq_0(X_0)\, d\gamma$$

Backward Process

- We know that

$$-\int (\nabla \cdot \mathbf{s})\, dq_t = -\iint (\nabla \cdot \mathbf{s})(e^{-t}X_0 + \sqrt{1-e^{-2t}}\gamma)\, dq_0(X_0)\, d\gamma$$

- Now, using Stein's formula ($\int \nabla \cdot \mathbf{v}\, d\gamma = \int \langle \gamma, \mathbf{v}(\gamma)\rangle\, d\gamma$, when γ is standard Gaussian.)

Statement of the lemma [edit]

Suppose X is a normally distributed random variable with expectation μ and variance σ^2. Further suppose g is a function for which the two expectations $E(g(X)(X-\mu))$ and $E(g'(X))$ both exist. (The existence of the expectation of any random variable is equivalent to the finiteness of the expectation of its absolute value.) Then

$$E(g(X)(X-\mu)) = \sigma^2 E(g'(X)).$$

Backward Process

- We know that

$$-\int (\nabla \cdot \mathbf{s})\, dq_t = -\iint (\nabla \cdot \mathbf{s})(e^{-t}X_0 + \sqrt{1-e^{-2t}}\gamma)\, dq_0(X_0)\, d\gamma$$

- Now, using Stein's formula ($\int \nabla \cdot \mathbf{v}\, d\gamma = \int \langle \gamma, \mathbf{v}(\gamma) \rangle\, d\gamma$, when γ is standard Gaussian.)

Statement of the lemma [edit]

Suppose X is a normally distributed random variable with expectation μ and variance σ^2. Further suppose g is a function for which the two expectations $E(g(X)(X-\mu))$ and $E(g'(X))$ both exist. (The existence of the expectation of any random variable is equivalent to the finiteness of the expectation of its absolute value.) Then

$$E(g(X)(X-\mu)) = \sigma^2 E(g'(X)).$$

- We know that:

$$-\iint (\nabla \cdot \mathbf{s})(e^{-t}X_0 + \sqrt{1-e^{-2t}}\gamma)\, d\gamma\, dq_0(X_0)$$

$$= -\frac{1}{\sqrt{1-e^{-2t}}} \int \langle \gamma, s(X_t) \rangle\, dq_0(X_0)\, d\gamma$$

Backward Process

- We know that:

$$-\iint (\nabla \cdot \mathbf{s})(e^{-t}X_0 + \sqrt{1-e^{-2t}}\gamma)\, dq_0(X_0)\, d\gamma$$

$$= -\frac{1}{\sqrt{1-e^{-2t}}} \int \langle \gamma, s(X_t) \rangle\, dq_0(X_0)\, d\gamma$$

Backward Process

- We know that:

$$-\iint (\nabla \cdot \mathbf{s})(e^{-t}X_0 + \sqrt{1-e^{-2t}}\gamma)\, dq_0(X_0)\, d\gamma$$

$$= -\frac{1}{\sqrt{1-e^{-2t}}} \int \langle \gamma, s(X_t) \rangle\, dq_0(X_0)\, d\gamma$$

- Which is:

$$= -\frac{1}{\sqrt{1-e^{-2t}}} \mathbb{E}_{X_t} \langle \gamma, s(X_t) \rangle$$

Backward Process

- Eventually, we know that we would like to minimize:

$$\min_s \mathbb{E}_{X_t}\left[\|s(X_t)\|^2 - 2\cdot\langle s(X_t), \nabla \ln q_t(X_t)\rangle + \|\nabla \ln q_t(X_t)\|^2\right]$$

Which is equivalent to:

$$= \min_s \mathbb{E}_{X_t}\left[\|s(X_t)\|^2 - 2\cdot\langle s(X_t), \nabla \ln q_t(X_t)\rangle\right]$$

Backward Process

- Eventually, we know that we would like to minimize:

$$\min_s \mathbb{E}_{X_t}\left[\|s(X_t)\|^2 - 2\cdot\langle s(X_t), \nabla \ln q_t(X_t)\rangle + \|\nabla \ln q_t(X_t)\|^2\right]$$

Which is equivalent to:

$$= \min_s \mathbb{E}_{X_t}\left[\|s(X_t)\|^2 - 2\cdot\langle s(X_t), \nabla \ln q_t(X_t)\rangle\right]$$

- Which is equal to:

$$= \min_s \mathbb{E}_{X_t}\|s(X_t)\|_2^2 - 2\left(-\frac{1}{\sqrt{1-e^{-2t}}}\mathbb{E}_{q_t}\langle\gamma, s(X_t)\rangle\right)$$

Backward Process

- Eventually, we know that we would like to minimize:

$$\min_s \mathbb{E}_{X_t}\left[\|s(X_t)\|^2 - 2\cdot\langle s(X_t), \nabla \ln q_t(X_t)\rangle + \|\nabla \ln q_t(X_t)\|^2\right]$$

  Which is equivalent to:

$$= \min_s \mathbb{E}_{X_t}\left[\|s(X_t)\|^2 - 2\cdot\langle s(X_t), \nabla \ln q_t(X_t)\rangle\right]$$

- Which is equal to:

$$= \min_s \mathbb{E}_{X_t}\|s(X_t)\|_2^2 - 2\left(-\frac{1}{\sqrt{1-e^{-2t}}}\mathbb{E}_{q_t}\langle\gamma, s(X_t)\rangle\right)$$

- Which is equivalent to :

$$\min_s \mathbb{E}_{X_t}\left\|s(X_t) + \frac{1}{\sqrt{1-e^{-2t}}}\gamma\right\|_2^2$$

Final Process

- This is the so-called "denoising diffusion probabilistic method (DDPM)"

Final Process

- This is the so-called "denoising diffusion probabilistic method (DDPM)"
- Final algorithm:

Final Process

- This is the so-called "denoising diffusion probabilistic method (DDPM)"
- Final algorithm:
- (1). Using data to estimate

$$\min_s \mathbb{E}_{X_t} \left\| s_t(X_t) + \frac{1}{\sqrt{1 - e^{-2t}}} \gamma \right\|_2^2$$

Where $X_0 \sim q^*$ is sampled from the target distribution.

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \gamma$$

Final Process

- This is the so-called "denoising diffusion probabilistic method (DDPM)"
- Final algorithm:
- (1). Using data to estimate

$$\min_s \mathbb{E}_{X_t} \left\| s_t(X_t) + \frac{1}{\sqrt{1 - e^{-2t}}}\gamma \right\|_2^2$$

Where $X_0 \sim q^*$ is sampled from the target distribution.

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}}\gamma$$

- (2). Then use the backward diffusion process to sample Y_T: starting from $Y_0 \sim \mathcal{N}(0, I)$ and

$$dY_t = Y_t + 2s_{T-t}(Y_t)dt + \sqrt{2}dB_t$$

Final Process

-