This Lecture: Advanced Machine Learning

- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Gaussian Processes
- Learning with Latent Variables
  - Probability Distributions
  - Approximate Inference
  - Mixture Models
  - EM and Generalizations
- Deep Learning
  - Neural Networks
  - CNNs, RNNs, RBMs, etc.

Topics of This Lecture

- Recap: Sampling approaches
  - Sampling from a distribution
  - Rejection Sampling
  - Importance Sampling
  - Sampling-Importance-Resampling
- Markov Chain Monte Carlo
  - Markov Chains
  - Metropolis Algorithm
  - Metropolis-Hastings Algorithm
  - Gibbs Sampling

Recap: Sampling Idea

- Objective:
  - Evaluate expectation of a function \( f(z) \) w.r.t. a probability distribution \( p(z) \).
  \[ \mathbb{E}[f] = \int f(z)p(z)\,dz \]
- Sampling idea:
  - Draw \( L \) independent samples \( z_l \) with \( l = 1, \ldots, L \) from \( p(z) \).
  - This allows the expectation to be approximated by a finite sum
  \[ \hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(z_l) \]
  - As long as the samples \( z_l \) are drawn independently from \( p(z) \), then
  \[ \mathbb{E}[\hat{f}] = \mathbb{E}[f] \]
  \( \Rightarrow \) Unbiased estimate, independent of the dimension of \( z \)!

Recap: Sampling from a pdf

- In general, assume we are given the pdf \( p(x) \) and the corresponding cumulative distribution:
  \[ F(x) = \int_{-\infty}^{x} p(z)\,dz \]
- To draw samples from this pdf, we can invert the cumulative distribution function:
  \[ u \sim \text{Uniform}(0,1) \Rightarrow F^{-1}(u) \sim p(x) \]

Note: Efficient Sampling from a Gaussian

- Problem with transformation method
  - Integral over Gaussian cannot be expressed in analytical form.
  - Standard transformation approach is very inefficient.
- More efficient: Box-Muller Algorithm
  - Generate pairs of uniformly distributed random numbers \( z_1, z_2 \in (-1,1) \).
  - Discard each pair unless it satisfies \( r^2 = z_1^2 + z_2^2 \leq 1 \).
  - This leads to a uniform distribution of points inside the unit circle with \( p(z_1, z_2) = 1/\pi \).
Box-Muller Algorithm (cont’d)

- Box-Muller Algorithm (cont’d)
  - For each pair $z_1, z_2$ evaluate
    $$y_1 = z_1 \left( \frac{-2 \ln r^2}{r^2} \right)^{1/2}, \quad y_2 = z_2 \left( \frac{-2 \ln r^2}{r^2} \right)^{1/2}$$
  - Then the joint distribution of $y_1$ and $y_2$ is given by
    $$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial (z_1, z_2)}{\partial (y_1, y_2)} \right|$$
    $$= \frac{1}{\sqrt{2\pi}} \exp\left(-y_1^2/2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-y_2^2/2\right)$$
    $$\Rightarrow y_1 \text{ and } y_2 \text{ are independent and each has a Gaussian distribution.}$$
- If $y \sim N(0,1)$, then $\sqrt{y} + \mu \sim N(\mu, \sigma^2)$.

Multivariate extension

- If $z$ is a vector valued random variable whose components are independent and Gaussian distributed with $N(0,1)$,
  - Then $y = \mu + Lz$ will have mean $\mu$ and covariance $\Sigma$.
  - Where $\Sigma = LL^T$ is the Cholesky decomposition of $\Sigma$.

General Advice

- Use library functions whenever possible
  - Many efficient algorithms available for known univariate distributions (and some other special cases)
  - This book (free online) explains how some of them work
    - http://www.nrbook.com/devroye/

Discussion

- Transformation method
  - Limited applicability, as we need to invert the indefinite integral of the required distribution $p(z)$.
  - This will only be feasible for a limited number of simple distributions.
- More general
  - Rejection Sampling
  - Importance Sampling

Rejection Sampling

- Assumptions
  - Sampling directly from $p(z)$ is difficult.
  - But we can easily evaluate $p(z)$ (up to some normalization factor $Z$):
    $$p(z) = \frac{1}{Z} p(z)$$
- Idea
  - We need some simpler distribution $q(z)$ (called proposal distribution) from which we can draw samples.
  - Choose a constant $k$ such that: $\forall z: kq(z) \geq \tilde{p}(z)$
- Sampling procedure
  - Generate a number $z_0$ from $q(z)$.
  - Generate a number $u_0$ from the uniform distribution over $[0, kq(z_0)]$.
  - If $u_0 > \tilde{p}(z_0)$ reject sample, otherwise accept.
    - Sample is rejected if it lies in the grey shaded area.
    - The remaining pairs $(u_0, z_0)$ have uniform distribution under the curve $\tilde{p}(z)$.
- Discussion
  - Original values of $z$ are generated from the distribution $q(z)$.
  - Samples are accepted with probability $\frac{\tilde{p}(z)}{kq(z)}$:
    $$p(\text{accept}) = \frac{\tilde{p}(z)}{kq(z)} = \frac{1}{k} \int \tilde{p}(z) dz$$
    $$\Rightarrow k \text{ should be as small as possible!}$$
Rejection Sampling - Discussion

- Limitation: high-dimensional spaces
  - For rejection sampling to be of practical value, we require that \( kq(z) \) be close to the required distribution, so that the rate of rejection is minimal.
- Artificial example
  - Assume that \( p(z) \) is Gaussian with covariance matrix \( \sigma_z^2 I \)
  - Assume that \( q(z) \) is Gaussian with covariance matrix \( \sigma_z^2 I \)
  - Obviously: \( \sigma_z^2 \geq \sigma_z^2 \)
  - In \( D \) dimensions: \( k = (\sigma_z/\sigma_z)^D \)
  - Assume \( \sigma_z \) is just 1% larger than \( \sigma_z \).
  - \( D = 1000 \Rightarrow k = 1.01^{1000} \geq 20,000 \)
  - And \( p(\text{accept}) = \frac{1}{20000} \)
  - Often impractical to find good proposal distributions for high dimensions.

Example: Sampling from a Gamma Distrib.

- Gamma distribution
  \[ \text{Gam}(z|a,b) = \frac{1}{\Gamma(a)} b^a z^{a-1} \exp(-bz) \quad a > 1 \]
- Rejection sampling approach
  - For \( a = 1 \), Gamma distribution has a bell-shaped form.
  - Suitable proposal distribution is Cauchy (for which we can use the transformation method).
  - Generalize Cauchy slightly to ensure it is nowhere smaller than Gamma: \( y = b \tan y + c \) for uniform \( y \).
  - This gives random numbers distributed according to \( q(z) = \frac{1}{1 + (z-c)^2/b^2} \) with optimal rejection rate for \( b^2 = 2a - 1 \).

Importance Sampling

- Idea
  - Method approximates expectations directly (but does not enable to draw samples from \( p(z) \) directly).
  - Use a proposal distribution \( q(z) \) from we can easily draw samples \( \tilde{z} \).
  - Express expectations in the form of a finite sum over samples \( \{\tilde{z}^{(l)}\} \) drawn from \( q(z) \).

Importance Sampling

- Removing the unknown normalization constants
  - We can use the sample set to evaluate the ratio of normalization constants
  \[ \frac{Z_p}{Z_q} = \frac{1}{Z_q} \int p(z)dz = \int \frac{p(z)}{q(z)} q(z)dz \approx \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_l q(\tilde{z}^{(l)}) \]
  - and therefore
  \[ \mathbb{E}[f] \approx \frac{1}{m} \sum_{l=1}^{L} w_l f(\tilde{z}^{(l)}) \]
  with \( w_l = \frac{\tilde{r}_l}{\sum_{m=1}^{m} \tilde{r}_m} \frac{p(\tilde{z}^{(l)})}{q(\tilde{z}^{(l)})} \)
  - In contrast to Rejection Sampling, all generated samples are retained (but they may get a small weight).
Importance Sampling - Discussion

• **Observations**
  - Success of Importance sampling depends crucially on how well the sampling distribution \(q(z)\) matches the desired distribution \(p(z)\).
  - Often, \(p(z)/q(z)\) is strongly varying and has a significant proportion of its mass concentrated over small regions of \(z\)-space.
  - Weights \(r\) may be dominated by a few weights having large values.
  - Practical issue: if none of the samples falls in the regions where \(p(z)/q(z)\) is large... The results may be arbitrary in error.
  - And there will be no diagnostic indication (no large variance in \(r\))!
  - Key requirement for sampling distribution \(q(z)\):
    - Should not be small or zero in regions where \(p(z)\) is significant!

Sampling-Importance-Resampling

• **Two stages**
  - Draw \(L\) samples \(z^{(1)},..., z^{(L)}\) from \(q(z)\).
  - Construct weights using importance weighting
    \[
    w_l = \frac{p(z^{(l)})}{q(z^{(l)})} \frac{q(z^{(l)})}{s(q(z^{(l)}))} 
    \]
    and draw a second set of samples \(z^{(1)},..., z^{(L)}\) with probabilities given by the weights \(w^{(1)},..., w^{(L)}\).

• **Result**
  - The resulting \(L\) samples are only approximately distributed according to \(p(z)\), but the distribution becomes correct in the limit \(L \to \infty\).

Sampling-Importance-Resampling (SIR)

• **Observation**
  - Success of rejection sampling depends on finding a good value for the constant \(k\).
  - For many pairs of distributions \(p(z)\) and \(q(z)\), it will be impractical to determine a suitable value for \(k\).
  - Any value that is sufficiently large to guarantee \(q(z) > p(z)\) will lead to impractically small acceptance rates.

• **Sampling-Importance-Resampling Approach**
  - Also makes use of a sampling distribution \(q(z)\), but avoids having to determine \(k\).

Curse of Dimensionality

• **Problem**
  - Rejection & Importance Sampling both scale badly with high dimensionality.
  - Example:
    \[
    p(z) \sim N(0, I), \quad q(z) \sim N(0, \sigma^2 I) 
    \]
  - Rejection Sampling
    - Requires \(\sigma \geq 1\). Fraction of proposals accepted: \(\sigma^{-\sigma}\).
  - Importance Sampling
    - Variance of importance weights:
      \[
      \frac{\sigma^2}{2 - 1/\sigma^2} - 1
      \]
    - Infinite / undefined variance if \(\sigma \leq 1/\sqrt{2}\).

Independent Sampling vs. Markov Chains

• **So far**
  - We’ve considered three methods, Rejection Sampling, Importance Sampling, and SIR, which were all based on independent samples from \(q(z)\).
  - However, for many problems of practical interest, it is often difficult or impossible to find \(q(z)\) with the necessary properties.
  - In addition, those methods suffer from severe limitations in high-dimensional spaces.

• **Different approach**
  - We abandon the idea of independent sampling.
  - Instead, rely on a Markov Chain to generate dependent samples from the target distribution.
  - Independence would be a nice thing, but it is not necessary for the Monte Carlo estimate to be valid.
MCMC - Markov Chain Monte Carlo

- **Overview**
  - Allows to sample from a large class of distributions.
  - Scales well with the dimensionality of the sample space.

- **Idea**
  - We maintain a record of the current state $z^{(t)}$.
  - The proposal distribution depends on the current state: $q(z_{A} | z_{B})$.
  - The sequence of samples forms a Markov chain $z^{(1)}, z^{(2)}, ...$

- **Setting**
  - We can evaluate $p(z)$ up to some normalizing factor $Z_p$.
  - At each time step, we generate a candidate sample from the proposal distribution and accept the sample according to a criterion.

MCMC – Metropolis Algorithm

- **Overview**
  - Proposal distribution is symmetric: $q(z_{A} | z_{B}) = q(z_{B} | z_{A})$.
  - The new candidate sample $z^{*}$ is accepted with probability $A(z^{*}, z^{(t)}) = \min\left(1, \frac{p(z^{*})}{p(z^{(t)})}\right)$.

- **Implementation**
  - Choose random number $u$ uniformly from unit interval $(0,1)$.
  - Accept sample if $A(z^{*}, z^{(t)}) > u$.

- **Note**
  - New candidate samples always accepted if $A(z^{*}, z^{(t)}) > 1$.
  - I.e., when new sample has higher probability than the previous one.

  This is in contrast to rejection sampling, where rejected samples are simply discarded.

  $\Rightarrow$ Leads to multiple copies of the same sample!

Line Fitting Example

- **Importance Sampling weights**
  - Many samples with very low weights...

Line Fitting Example (cont’d)

- **Metropolis algorithm**
  - Perturb parameters: $Q(z' | z)$, e.g. $N(z, \sigma^2)$.
  - Accept with probability $\min\left(1, \frac{p(z')}{p(z)}\right)$.
  - Otherwise, keep old parameters.
Markov Chains

- **Question**
  - How can we show that \( z' \) tends to \( p(x) \) as \( t \to \infty \)?

- **Markov chains**
  - First-order Markov chain:
    \[ p(z^{(m+1)} | z^{(m)}, \ldots, z^{(1)}) = p(z^{(m+1)} | z^{(m)}) \]
  - Marginal probability
    \[ p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)} | z^{(m)}) p(z^{(m)}) \]
  - A Markov chain is called homogeneous if the transition probabilities \( p(z^{(m+1)} | z^{(m)}) \) are the same for all \( m \).

- **Invariant distribution**
  - A distribution is said to be invariant (or stationary) w.r.t. a Markov chain if each step in the chain leaves that distribution invariant.
  - Transition probabilities:
    \[ T(z^{(m)}, z^{(m+1)}) = p(z^{(m+1)} | z^{(m)}) \]
  - For homogeneous Markov chain, distribution \( p'(z) \) is invariant if:
    \[ p'(z) = \sum_{z'} T(z', z) p'(z') \]
  - Detailed balance
    - Sufficient (but not necessary) condition to ensure that a distribution is invariant:
      \[ p'(z) T(z, z') = p'(z') T(z', z) \]
    - A Markov chain which respects detailed balance is reversible.

Detailed Balance

- **Detailed balance** means
  - If we pick a state from the target distribution \( p(x) \) and make a transition under \( T \) to another state, it is just as likely that we will pick \( x_j \) and go from \( x_j \) to \( x_k \) than that we will pick \( x_k \) and go from \( x_k \) to \( x_j \).
  - It can easily be seen that a transition probability that satisfies detailed balance w.r.t. a particular distribution will leave that distribution invariant, because
    \[ \sum_{z'} p'(z') T(z', z) = \sum_{z'} p'(z) T(z, z') = p'(z) \sum_{z'} p'(z') = p'(z) \]

Mixture Transition Distributions

- **Mixture distributions**
  - In practice, we often construct the transition probabilities from a set of "base" transitions \( B_1, \ldots, B_k \).
  - This can be achieved through a mixture distribution
    \[ T(z'; z) = \sum_{j=1}^k \alpha_j B_j(z'; z) \]
    with mixing coefficients \( \alpha_j \geq 0 \) and \( \sum \alpha_j = 1 \).

- **Properties**
  - If the distribution is invariant w.r.t. each of the base transitions, then it will also be invariant w.r.t. \( T(z'; z) \).
  - If each of the base transitions satisfies detailed balance, then the mixture transition \( T \) will also satisfy detailed balance.

  Common example: each base transition changes only a subset of variables.

MCMC - Metropolis-Hastings Algorithm

- **Metropolis-Hastings Algorithm**
  - Generalization: Proposal distribution not required to be symmetric.
  - The new candidate sample \( z' \) is accepted with probability
    \[ A(z', z) = \min \left( 1, \frac{p(z') q(z | z')}{p(z) q(z' | z)} \right) \]
    where \( k \) labels the members of the set of possible transitions considered.

- **Note**
  - Evaluation of acceptance criterion does not require normalizing constant \( Z_p \).
  - When the proposal distributions are symmetric, Metropolis-Hastings reduces to the standard Metropolis algorithm.
MCMC - Metropolis-Hastings Algorithm

- **Properties**
  - We can show that \( p(z) \) is an invariant distribution of the Markov chain defined by the Metropolis-Hastings algorithm.
  - We show detailed balance:
    \[
    A(z', z) = \min \left\{ \frac{1}{A(z, z')} \left| \frac{p(z')q_0(z'|z)}{p(z)q_0(z|z')} \right| \right\}
    \]
    \[
    \pi(z)q_0(z'|z)A(z, z') = \min \left\{ \frac{1}{A(z, z')} \left| \frac{p(z')q_0(z'|z)}{p(z)q_0(z|z')} \right| \right\}
    \]
    \[
    \pi(z)q_0(z'|z)A(z, z') = \pi(z')q_0(z'|z)A(z', z)
    \]
    \[
    \pi(z)q_0(z'|z)A(z, z') = \pi(z')q_0(z'|z)A(z', z)
    \]
  - Central goal in MCMC is to avoid random walk behavior!

  *Note: This is wrong in the Bishop book!*

Random Walks

- **Example**: Random Walk behavior
  - Consider a state space consisting of the integers \( z \in \mathbb{Z} \) with initial state \( z(1) = 0 \) and transition probabilities
    \[
    p(z(t+1) = z(t) + 1) = 0.25
    \]
    \[
    p(z(t+1) = z(t) - 1) = 0.25
    \]

- **Analysis**
  - Expected state at time \( t \):
    \[
    \mathbb{E}[z(t)] = 0
    \]
  - Variance:
    \[
    \mathbb{E}[(z(t))^2] = t/2
    \]
  - After \( t \) steps, the random walk has only traversed a distance that is on average proportional to \( \sqrt{t} \).
  - Central goal in MCMC is to avoid random walk behavior!

MCMC - Metropolis-Hastings Algorithm

- **Schematic illustration**
  - For continuous state spaces, a common choice of proposal distribution is a Gaussian centered on the current state.

  *What should be the variance of the proposal distribution?*
  - Large variance: rejection rate will be high for complex problems.
  - The scale \( \sigma \) of the proposal distribution should be as large as possible without incurring high rejection rates.
  - \( \sigma \) should be of the same order as the smallest length scale \( \eta_{min} \).
  - This causes the system to explore the distribution by means of a random walk.

  *Undesired behavior: number of steps to arrive at state that is independent of original state is of order \( (\sigma^2/\eta_{min})^2 \).*
  - Strong correlations can slow down the Metropolis-Hastings algorithm!

Gibbs Sampling

- **Approach**
  - MCMC-algorithm that is simple and widely applicable.

- **May be seen as a special case of Metropolis-Hastings.**

- **Idea**
  - Sample variable-wise: replace \( z_i \) by a value drawn from the distribution \( p(z_i|z_{\bar{i}}) \).
  - This means we update one coordinate at a time.
  - Repeat procedure either by cycling through all variables or by choosing the next variable.

Gibbs Sampling

- **Example**
  - Assume distribution \( p(z_1, z_2, z_3) \).
  - Replace \( z_1 \) with new value drawn from \( p(z_1|z_2, z_3) \).
  - Replace \( z_2 \) with new value drawn from \( p(z_2|z_1, z_3) \).
  - Replace \( z_3 \) with new value drawn from \( p(z_3|z_1, z_2) \).
  - And so on...
Discussion

- Gibbs sampling benefits from few free choices and convenient features of conditional distributions:
  - Conditionals with a few discrete settings can be explicitly normalized:
    \[ p(x_i|x_{j \neq i}) = \frac{p(x_i, x_{j \neq i})}{\sum_x p(x_i, x_{j \neq i})} \]
    This sum is small and easy.
  - Continuous conditionals are often only univariate.
    \[ \Rightarrow \text{amenable to standard sampling methods.} \]
  - In case of graphical models, the conditional distributions depend only on the variables in the corresponding Markov blankets.

Gibbs Sampling

- Example
  - 20 iterations of Gibbs sampling on a bivariate Gaussian.

  \[ \text{Note: strong correlations can slow down Gibbs sampling.} \]

How Should We Run MCMC?

- Arbitrary initialization means starting iterations are bad
  - Discard a "burn-in" period.
- How do we know if we have run for long enough?
  - You don’t. That’s the problem.
- The samples are not independent
  - Solution 1: Keep only every \( M \)th sample ("thinning").
  - Solution 2: Keep all samples and use the simple Monte Carlo estimator on MCMC samples
    \[ \text{It is consistent and unbiased if the chain has "burned in".} \]
    \[ \Rightarrow \text{Use thinning only if computing } f(x) \text{ is expensive.} \]
- For opinion on thinning, multiple runs, burn in, etc.

Summary: Approximate Inference

- Exact Bayesian Inference often intractable.
- Rejection and Importance Sampling
  - Generate independent samples.
  - Impractical in high-dimensional state spaces.
- Markov Chain Monte Carlo (MCMC)
  - Simple & effective (even though typically computationally expensive).
  - Scales well with the dimensionality of the state space.
  - Issues of convergence have to be considered carefully.
- Gibbs Sampling
  - Used extensively in practice.
  - Parameter Free
  - Requires sampling conditional distributions.

References and Further Reading

- Sampling methods for approximate inference are described in detail in Chapter 11 of Bishop’s book.
- Another good introduction to Monte Carlo methods can be found in Chapter 29 of MacKay’s book (also available online: http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.html)