# Sampling

Frank Wood

Many thanks to Iain Murray (University of Toronto / University of Edinburgh) from whom most of these slides were borrowed with permission. Many of these slides appear in his machine learning summer school tutorial from 2009

November 24, 2009

# Bayesian Inference In Regression

Regression can be expressed in the Bayesian framework where the goal is to learn the posterior distribution over the model parameters (slope and intercept).

$$P(\mathbf{w}|\mathbf{t}) \quad \propto \quad P(\mathbf{t}|\mathbf{w})P(\mathbf{w})$$

Then, if we know $P(\mathbf{w}|\mathbf{t})$ then we can compute or analytically derive the posterior predictive distribution

$$P(t|\mathbf{t}, \beta, \lambda) = \int P(t|\mathbf{w}, \beta)P(\mathbf{w}|\mathbf{t}, \beta, \lambda)d\mathbf{w}$$

In the Gaussian linear regression case we can (we'll derive that result later in the class). For now, let's develop some intuition about what's going on.

# Discrete Prior

For pedagogical purposes, let us consider a discrete prior

$$\{\mathbf{w}_\ell\}_{\ell=1}^{L}, \mathbf{w}_\ell \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$$

This can be expressed as a uniformly weighted mixture

$$P(\mathbf{w}) = \frac{1}{L}\sum_{\ell} \delta_{\mathbf{w}_\ell}$$

# Posterior Estimation, Discrete Prior

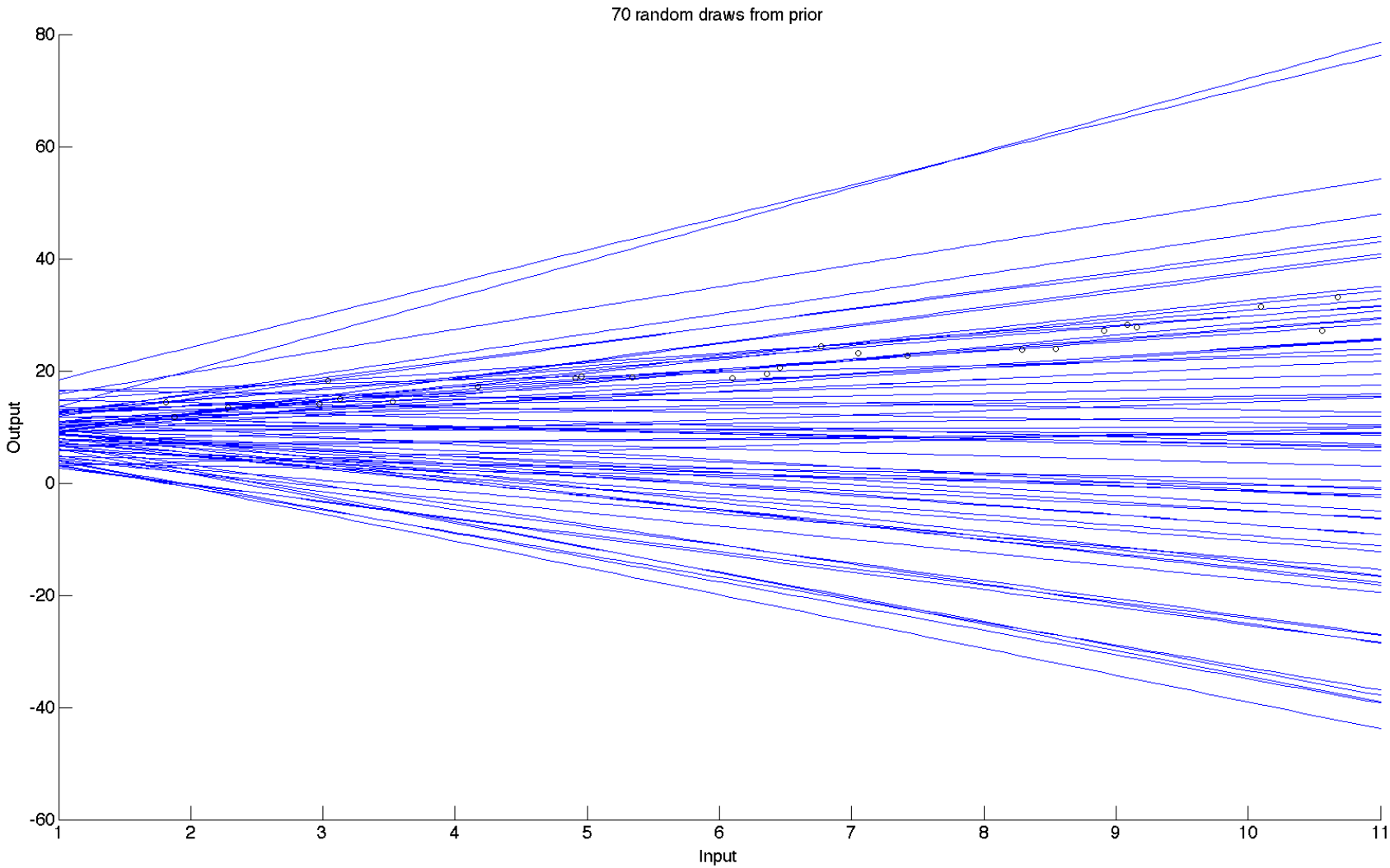



Figure: Random draws from the prior

# Discrete Posterior

Given a discrete prior, the posterior can likewise be expressed as a (non-uniformly) weighted discrete mixture

$$P(\mathbf{w}|\mathbf{t}) \propto \int P(\mathbf{t}|\mathbf{w})\delta(\mathbf{w} - \mathbf{w}_\ell)d\mathbf{w} = \sum_\ell P(\mathbf{t}|\mathbf{w}_\ell)$$

Where the weights for the discrete posterior are proportional to the likelihood of the data under a regression model with corresponding parameter

$$P(\mathbf{w}|\mathbf{t}) = \sum_\ell \pi_\ell \delta_{\mathbf{w}_\ell}, \pi_\ell = \frac{P(\mathbf{t}|\mathbf{w}_\ell)}{\sum_j P(\mathbf{t}|\mathbf{w}_j)}$$

# Weighted Posterior Distribution Over Model Parameters

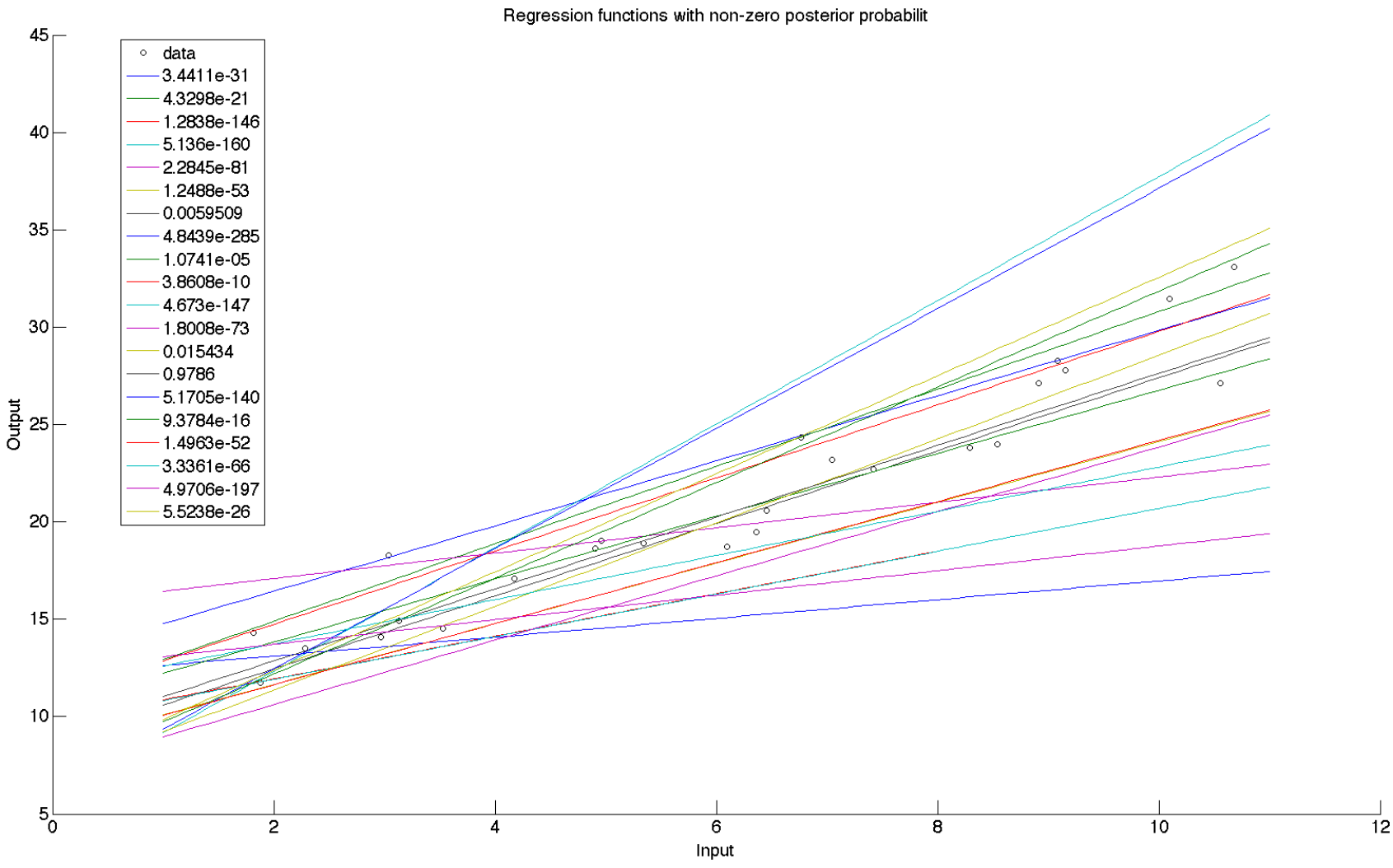



Figure: Weight vectors with high posterior probability.

# Posterior Predictive Calculation with Discrete Posterior

Given the simple form of a discrete posterior distribution, the posterior predictive distribution can also be straightforwardly calculated.

$$P(t|\mathbf{t}, \beta, \lambda) = \int P(t|\mathbf{w}, \beta) P(\mathbf{w}|\mathbf{t}, \beta, \lambda) d\mathbf{w}$$

Becomes

$$P(t|\mathbf{t}, \beta, \lambda) = \sum_{\ell} \pi_{\ell} P(t|\mathbf{w}_{\ell}, \beta)$$

# Weighted Posterior Distribution Over Model Parameters

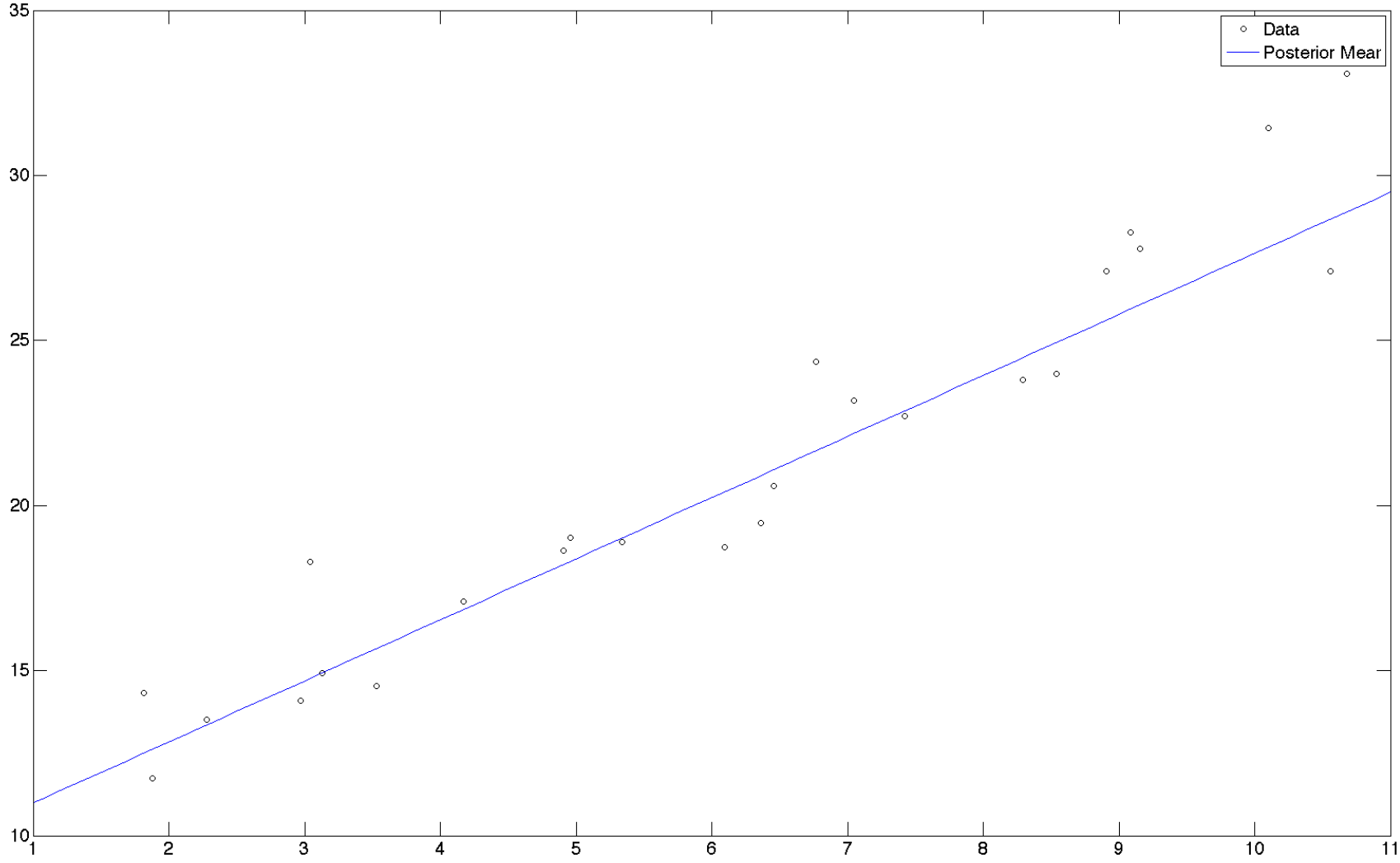



Figure: Posterior predictive distribution.

# Bayesian Regression

But this prior is not very smart.

We've seen how one choice of noninformative prior gives rise to analytic posterior and posterior predictive distributions.

What about different choices of prior? How do we do inference if the posterior distribution isn't tractable?

*Answer: Sampling*

# A statistical problem

**What is the average height of the the people in this class ?**
Method: measure their heights, add them up and divide by $N = 20$.

**What is the average height $f$ of people $p$ in New York $\mathcal{C}$?**

$$E_{p \in \mathcal{C}}[f(p)] \equiv \frac{1}{|\mathcal{C}|} \sum_{p \in \mathcal{C}} f(p), \quad \text{“intractable”?}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} f\left(p^{(s)}\right), \quad \text{for random survey of } S \text{ people } \{p^{(s)}\} \in \mathcal{C}$$

Surveying works for large and notionally infinite populations.

# Simple Monte Carlo

Statistical sampling can be applied to any expectation:

**In general:**

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

**Example: making predictions**

$$p(x|\mathcal{D}) = \int P(x|\theta,\mathcal{D})P(\theta|\mathcal{D})\,\mathrm{d}\theta$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} P(x|\theta^{(s)},\mathcal{D}), \quad \theta^{(s)} \sim P(\theta|\mathcal{D})$$

# Properties of Monte Carlo

Estimator: $\int f(x)P(x)\,\mathrm{d}x \approx \hat{f} \equiv \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$

**Estimator is unbiased:**

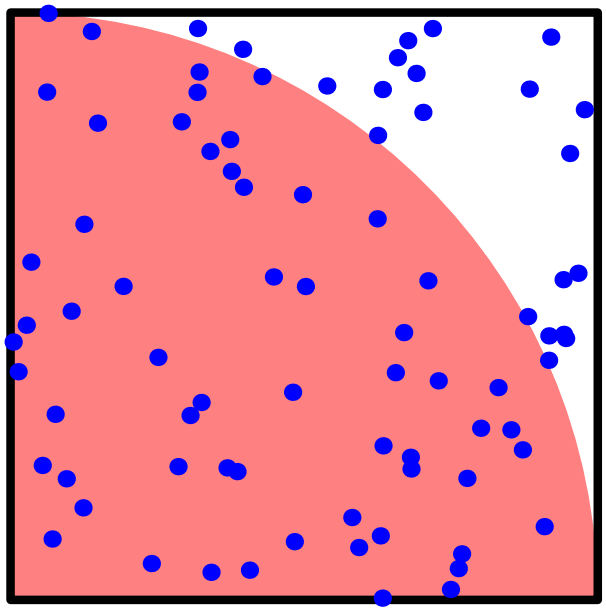
$$\mathbb{E}_{P(\{x^{(s)}\})}\left[\hat{f}\right] = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{P(x)}[f(x)] = \mathbb{E}_{P(x)}[f(x)]$$

**Variance shrinks $\propto 1/S$:**

$$\mathrm{var}_{P(\{x^{(s)}\})}\left[\hat{f}\right] = \frac{1}{S^2}\sum_{s=1}^{S}\mathrm{var}_{P(x)}[f(x)] = \mathrm{var}_{P(x)}[f(x)]\,/S$$

"Error bars" shrink like $\sqrt{S}$

# A dumb approximation of $\pi$



$$P(x,y) = \begin{cases} 1 & 0<x<1 \quad \text{and} \quad 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\, \mathrm{d}x\, \mathrm{d}y$$

# Sampling from distributions

How to convert samples from a Uniform[0,1] generator:

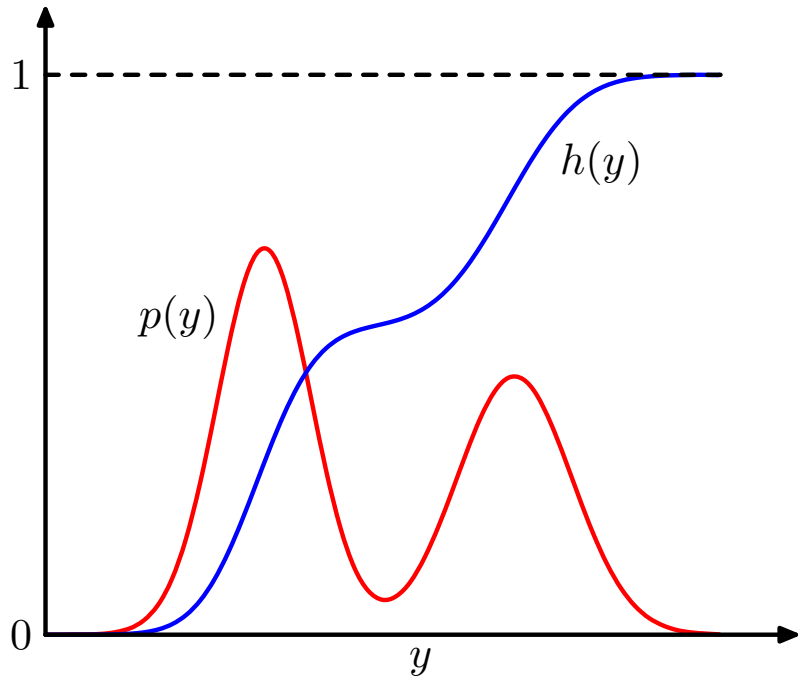



Figure from PRML, Bishop (2006)

$$h(y) = \int_{-\infty}^{y} p(y') \, \mathrm{d}y'$$

Draw mass to left of point:

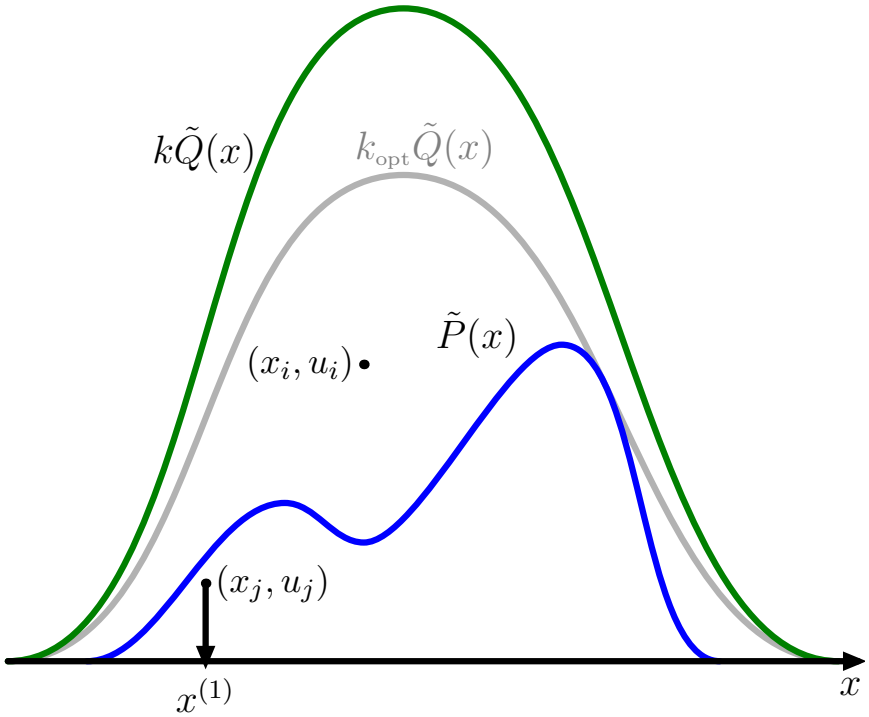
$$u \sim \text{Uniform}[0,1]$$

Sample, $y(u) = h^{-1}(u)$

Although we can't always compute and invert $h(y)$

# Rejection sampling

Sampling underneath a $\tilde{P}(x) \propto P(x)$ curve is also valid




Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$:

- Draw $x \sim Q(x)$
- height $u \sim \text{Uniform}[0, k\tilde{Q}(x)]$

Discard the point if above $\tilde{P}$, i.e. if $u > \tilde{P}(x)$

# Importance sampling

Computing $\tilde{P}(x)$ and $\tilde{Q}(x)$, then *throwing x away* seems wasteful
Instead rewrite the integral as an expectation under $Q$:

$$\int f(x)P(x)\,\mathrm{d}x = \int f(x)\frac{P(x)}{Q(x)}Q(x)\,\mathrm{d}x, \qquad (Q(x) > 0 \text{ if } P(x) > 0)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)} \sim Q(x)$$

This is just simple Monte Carlo again, so it is unbiased.

Importance sampling applies when the integral is not an expectation.
Divide and multiply any integrand by a convenient distribution.

# Importance sampling (2)

Previous slide assumed we could evaluate $P(x) = \tilde{P}(x)/\mathcal{Z}_P$

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{\mathcal{Z}_Q}{\mathcal{Z}_P}\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\underbrace{\frac{\tilde{P}(x^{(s)})}{\tilde{Q}(x^{(s)})}}_{\tilde{r}^{(s)}}, \quad x^{(s)} \sim Q(x)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{\tilde{r}^{(s)}}{\frac{1}{S}\sum_{s'}\tilde{r}^{(s')}} \equiv \sum_{s=1}^{S} f(x^{(s)})w^{(s)}$$

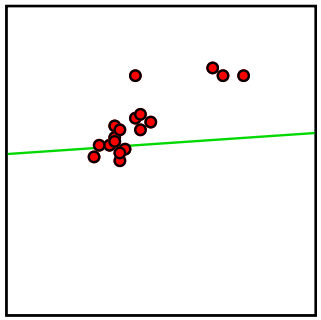
This estimator is **consistent** but **biased**

**Exercise:** Prove that $\mathcal{Z}_P/\mathcal{Z}_Q \approx \frac{1}{S}\sum_s \tilde{r}^{(s)}$

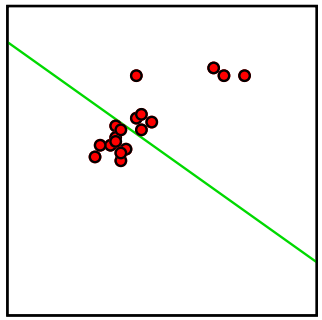
# Summary so far

- Sums and integrals, often expectations, occur frequently in statistics
- **Monte Carlo** approximates expectations with a sample average
- **Rejection sampling** draws samples from complex distributions
- **Importance sampling** applies Monte Carlo to 'any' sum/integral

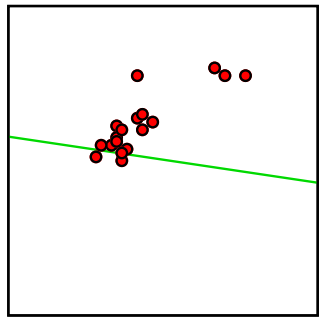
# Importance sampling weights



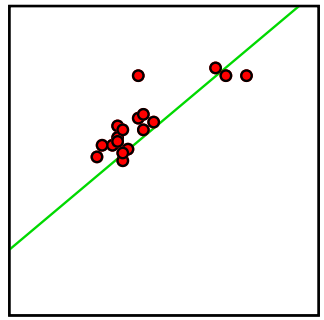
$w$ =0.00548



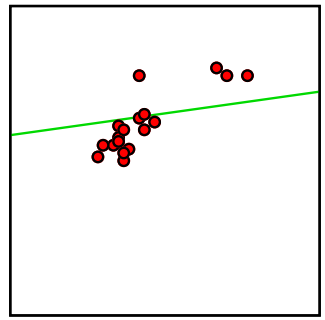
$w$ =1.59e-08



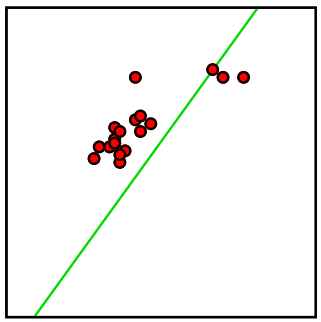
$w$ =9.65e-06



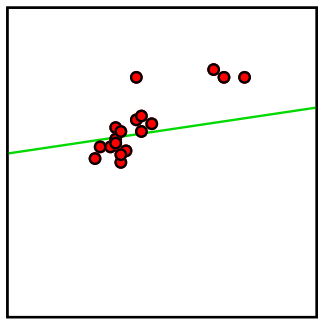
$w$ =0.371



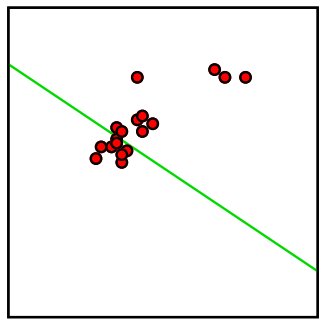
$w$ =0.103



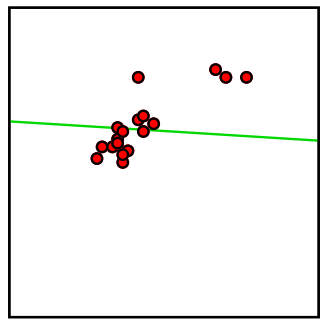
$w$ =1.01e-08



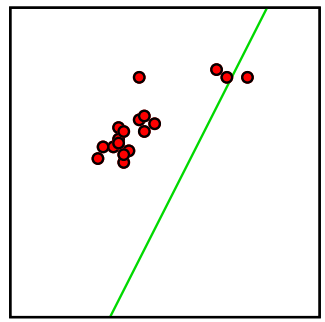
$w$ =0.111



$w$ =1.92e-09

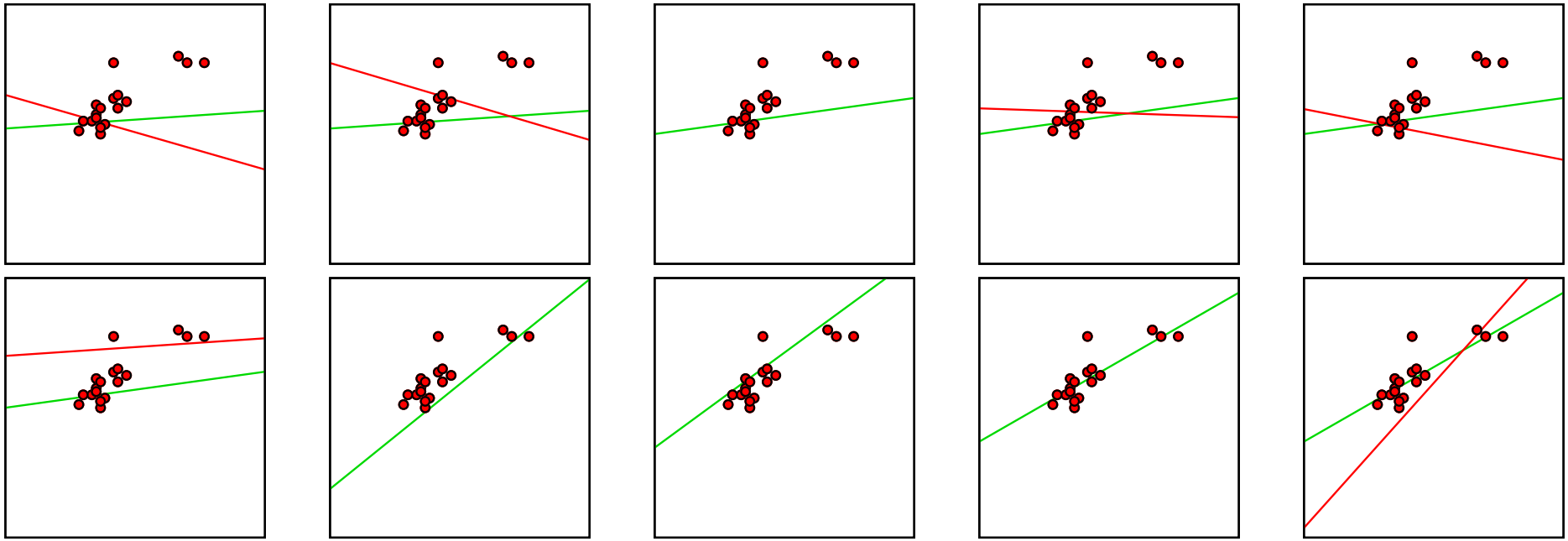


$w$ =0.0126

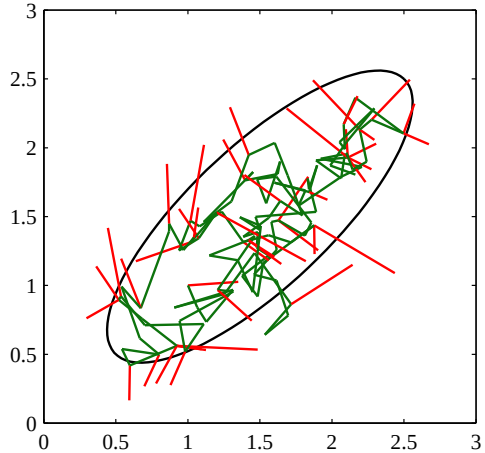


$w$ =1.1e-51

# Metropolis algorithm



- Perturb parameters: $Q(\theta'; \theta)$, e.g. $\mathcal{N}(\theta, \sigma^2)$
- Accept with probability $$\min\left(1, \frac{\tilde{P}(\theta'|\mathcal{D})}{\tilde{P}(\theta|\mathcal{D})}\right)$$
- Otherwise **keep old parameters**



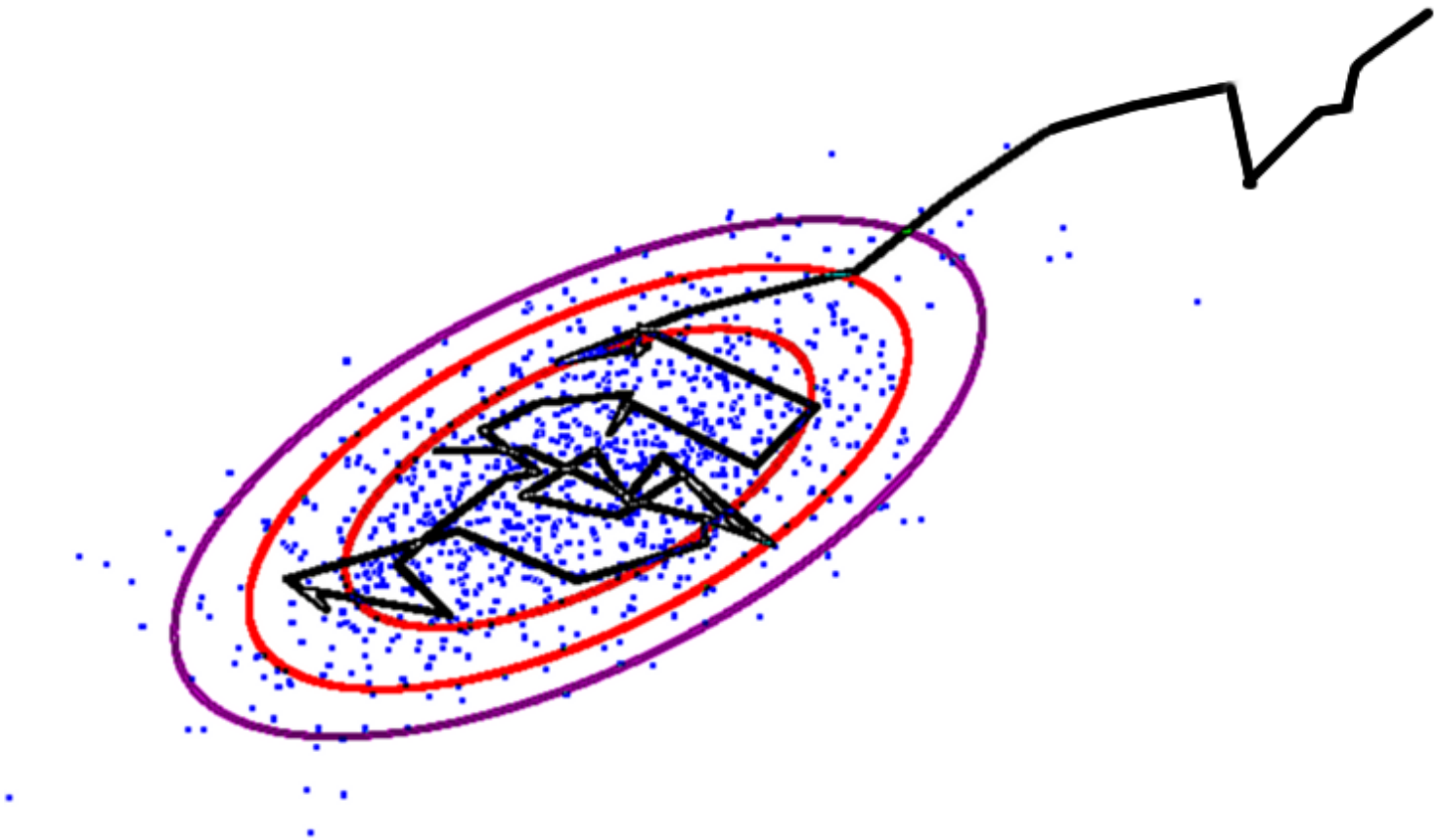

This subfigure from Bishop (2006)

Detail: Metropolis, as stated, requires $Q(\theta'; \theta) = Q(\theta; \theta')$

# Markov chain Monte Carlo

**Construct a biased random walk that explores target dist $P^\star(x)$**

Markov steps, $x_t \sim T(x_t \leftarrow x_{t-1})$



MCMC gives approximate, correlated samples from $P^\star(x)$

# Transition operators

## Discrete example

$$P^\star = \begin{pmatrix} 3/5 \\ 1/5 \\ 1/5 \end{pmatrix} \qquad T = \begin{pmatrix} 2/3 & 1/2 & 1/2 \\ 1/6 & 0 & 1/2 \\ 1/6 & 1/2 & 0 \end{pmatrix} \qquad T_{ij} = T(x_i \leftarrow x_j)$$

$P^\star$ is an **invariant distribution** of $T$ because $TP^\star = P^\star$, i.e.

$$\sum_x T(x' \leftarrow x) P^\star(x) = P^\star(x')$$

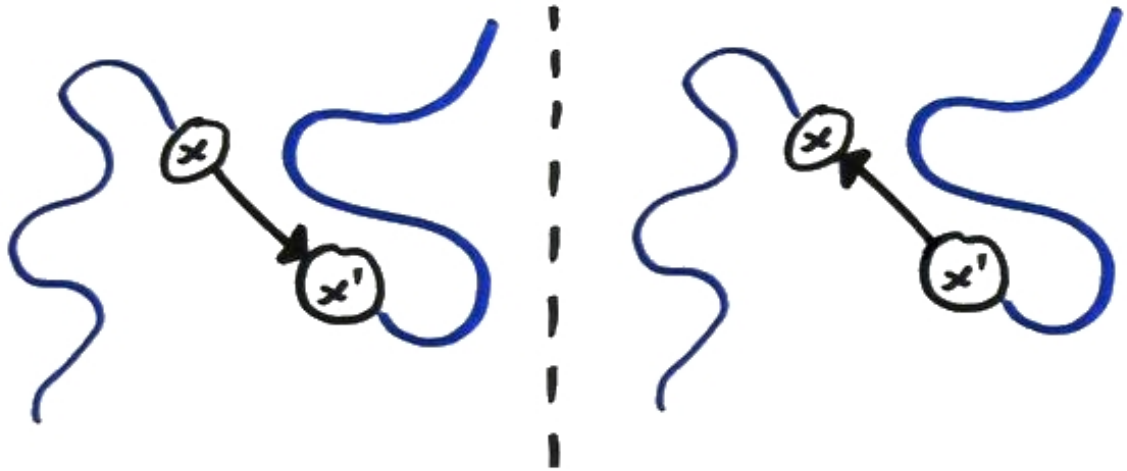
Also $P^\star$ is *the* **equilibrium distribution** of $T$:

To machine precision: $T^{100} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3/5 \\ 1/5 \\ 1/5 \end{pmatrix} = P^\star$

*Ergodicity* requires: $T^K(x' \leftarrow x) > 0$ for all $x' : P^\star(x') > 0$, for some $K$

# Detailed Balance

Detailed balance means $\rightarrow x \rightarrow x'$ and $\rightarrow x' \rightarrow x$ are equally probable:




$$T(x' \leftarrow x)P^\star(x) = T(x \leftarrow x')P^\star(x')$$

**Detailed balance implies the invariant condition:**

$$\sum_x T(x' \leftarrow x)P^\star(x) = P^\star(x')\sum_x T(x \leftarrow x') = 1$$

Enforcing detailed balance is easy: it only involves isolated pairs

# Reverse operators

If $T$ satisfies stationarity, we can define a *reverse operator*

$$\begin{aligned}\widetilde{T}(x \leftarrow x') &\propto T(x' \leftarrow x)\, P^\star(x) \\ &= \frac{T(x' \leftarrow x)\, P^\star(x)}{\sum_x T(x' \leftarrow x)\, P^\star(x)} = \frac{T(x' \leftarrow x)\, P^\star(x)}{P^\star(x')}.\end{aligned}$$

**Generalized balance condition:**

$$T(x' \leftarrow x) P^\star(x) \;=\; \widetilde{T}(x \leftarrow x') P^\star(x')$$

also implies the invariant condition *and is necessary*.

Operators satisfying detailed balance are their own reverse operator.

# Metropolis–Hastings

## Transition operator

- Propose a move from the current state $Q(x'; x)$, e.g. $\mathcal{N}(x, \sigma^2)$
- Accept with probability $\min\left(1, \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right)$
- Otherwise next state in chain is a copy of current state

## Notes

- Can use $\tilde{P} \propto P(x)$; normalizer cancels in acceptance ratio
- Satisfies detailed balance (shown below)
- $Q$ must be chosen to fulfill the other technical requirements

---

$$P(x) \cdot T(x' \leftarrow x) = P(x) \cdot Q(x'; x) \min\left(1, \frac{P(x')Q(x; x')}{P(x)Q(x'; x)}\right) = \min\Big(P(x)Q(x'; x),\ P(x')Q(x; x')\Big)$$

$$= P(x') \cdot Q(x; x') \min\left(1, \frac{P(x)Q(x'; x)}{P(x')Q(x; x')}\right) =$$

$$P(x') \cdot T(x \leftarrow x')$$