# Sampling by Optimization

Jan-Hendrik Lange, Riad Akrour

July 16, 2015

# Outline

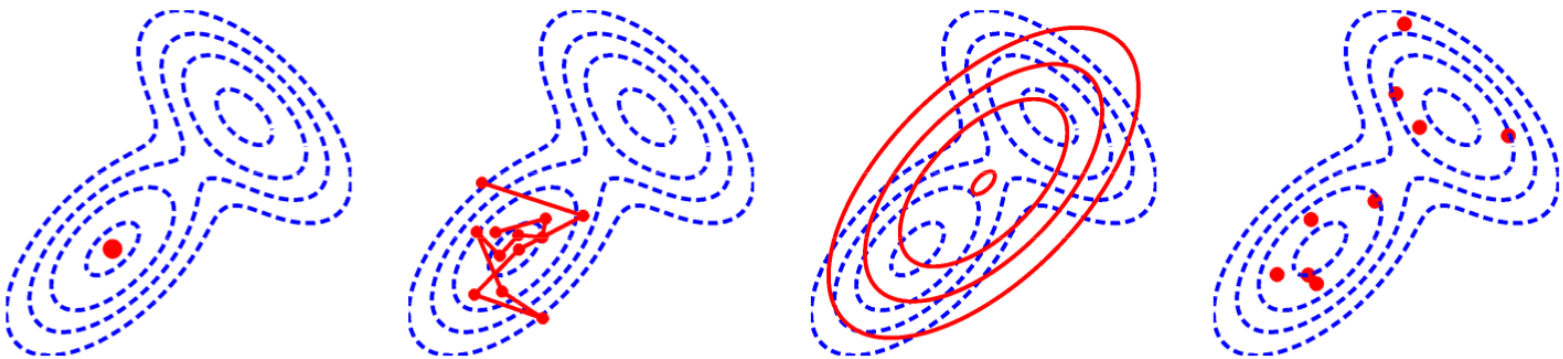
**We discuss the following topics:**

1. Approximate samples from Gaussian MRFs
   (Papandreou, Yuille, NIPS 2010)
2. Approximate samples from discrete Gibbs MRFs
   (Papandreou, Yuille, ICCV 2011)
3. Bounds on the approximation
   (Hazan, Jaakkola, ICML 2012)
4. Exact samples from continuous distributions
   (Maddison, Tarlow, Minka, NIPS 2014)

# Why do we need to sample?

- Computing expectations of stochastic models
- Generate content from models for applications such as
  - inpainting
  - interaction with a user

# Sampling Techniques



Left to right: MAP, MCMC, Variational Bayes, Perturbation & Maximization

Sampling by Optimization: Sample from some model $p(x \mid \theta)$ via the following steps:

a) Perturb the model to $p(x \mid \theta + \epsilon)$.

b) Find $x^* = \arg\max p(x \mid \theta + \epsilon)$.

Approximate samples from Gaussian MRFs
(Papandreou, Yuille, NIPS 2010)

# Efficient sampling from Gaussian Markov Random Fields

## Definition

Let a density function with parameters $\{F, \mu_0, \Sigma_0\}$ be defined by:

$$\begin{aligned} p(x) &\propto \mathcal{N}(Fx|\mu_0, \Sigma_0) \\ &\propto \exp(-\frac{1}{2}(Fx-\mu_0)^T\Sigma_0^{-1}(Fx-\mu_0)) \\ &= \mathcal{N}(J^{-1}k, J^{-1}) \end{aligned}$$

With $J = F^T\Sigma_0^{-1}F$ and $k = F^T\Sigma_0^{-1}\mu_0$

## Naive Sampling

With Cholesky decomposition of $J^{-1} \Rightarrow$ at least $\mathcal{O}(N^2)$ complexity

## Sampling by optimization

Perturb parameters and find approximate MAP $\Rightarrow \mathcal{O}(N^{\frac{3}{2}})$ complexity

# Efficient sampling from Gaussian Markov Random Fields

## Recall: Gaussian MRF

$x \sim \mathcal{N}(J^{-1}k, J^{-1})$ with $J = F^T\Sigma_0^{-1}F$ and $k = F^T\Sigma_0^{-1}\mu_0$

## Exactness

For $\tilde{\mu}_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $\tilde{x}$ MAP of $\mathcal{N}(Fx|\tilde{\mu}_0, \Sigma_0) \Rightarrow \tilde{x} \sim \mathcal{N}(J^{-1}k, J^{-1})$
Proof:

$$\tilde{x} \stackrel{\Delta}{=} \arg\max_x \mathcal{N}(Fx|\tilde{\mu}_0, \Sigma_0)$$

$$\tilde{x} = J^{-1}F^T\Sigma_0^{-1}\tilde{\mu}_0 \qquad \text{(the mean)}$$

$$\tilde{x} \sim \mathcal{N}(J^{-1}k, J^{-1}F^T\Sigma_0^{-1}FJ^{-1}) \qquad \text{(Affine transformation)}$$

$$\tilde{x} \sim \mathcal{N}(J^{-1}k, J^{-1})$$

## The trick

$\tilde{x}$ solution of $J\tilde{x} = B$. Instead of inverting $J$, use approximate methods (e.g. multigrid) to find $\tilde{x}$ more efficiently

# Common questions

- To which parameters should the noise be added?
- Which kind of noise should be added?
- What's the approximation error of the overall algorithm?

Approximate samples from discrete Gibbs MRFs
(Papandreou, Yuille, ICCV 2011)

# Sampling by Optimization for Gibbs distributions

Consider the general (finite) Gibbs distribution, with potential function $\phi$, where for any state $x \in X$ we have:

$$p(x) \propto \exp(\phi(x)).$$

We add to each potential $\phi(x)$ a random perturbation $\epsilon_x$ to obtain the new (perturbed) potential

$$\gamma(x) = \phi(x) + \epsilon_x.$$

The search for suitable perturbations leads to the *Gumbel* distribution.

# Gumbel Distribution

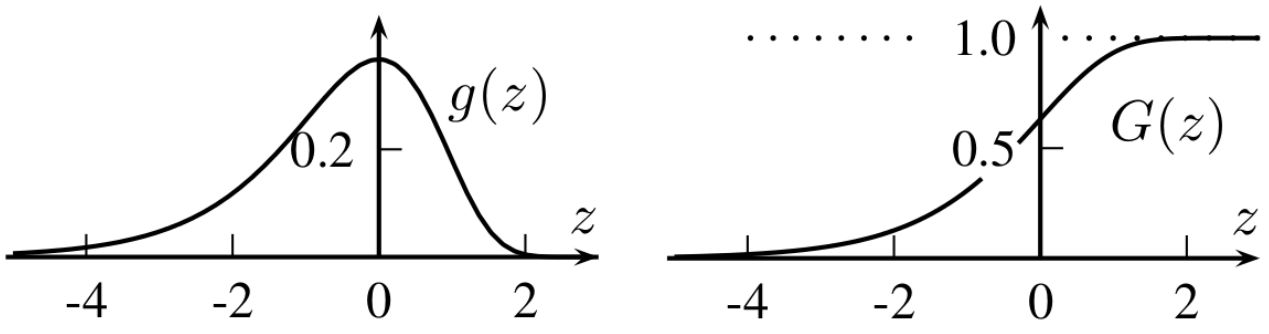
A continuous univariate random variable is $\mathrm{Gumbel}(a)$ (Gumbel with location $a$), if it has log-concave density

$$g(z) = \exp(-(z-a) + e^{-(z-a)}).$$

Thus it can be efficiently sampled by applying the inverse of its CDF

$$G(z) = \exp(-e^{-(z-a)})$$

to standard uniform samples.




$\mathrm{Gumbel}(0)$: PDF and CDF

# Gumbel Perturbations

Let $\epsilon_x$ be random i.i.d. $\mathrm{Gumbel}(0)$ samples. Then

$$\gamma(x) = \phi(x) + \epsilon_x \sim \mathrm{Gumbel}(\phi(x)),$$

i.e. the perturbed potential is Gumbel with location $\phi(x)$.
Moreover, we have the identity

$$\mathrm{P}[\arg\max_x \gamma(x) = \bar{x}] = \frac{\exp(\phi(\bar{x}))}{\sum_x \exp(\phi(x))}.$$

# Proof of the distribution of the $\arg\max$

By definition of $\gamma$ we have

$$\begin{aligned}
\mathrm{P}[\arg\max_x \gamma(x) = \bar{x}] &= \mathrm{P}[\gamma(\bar{x}) \geq \max_{x \neq \bar{x}} \gamma(x)] \\
&= \int_{-\infty}^{\infty} g(t; \phi(\bar{x})) \prod_{x \neq \bar{x}} G(t; \phi(x)) \, \mathrm{d}t \\
&= \int_{-\infty}^{\infty} e^{\phi(\bar{x}) - t} \exp(-e^{\phi(\bar{x}) - t}) \prod_{x \neq \bar{x}} \exp(-e^{\phi(x) - t}) \, \mathrm{d}t.
\end{aligned}$$

# Proof of the distribution of the $\arg\max$

We substitute $z = \exp(-e^{\phi(\bar{x})-t})$. This yields

$$\mathrm{d}z = e^{\phi(\bar{x})-t} \exp(-e^{\phi(\bar{x})-t}) \, \mathrm{d}t$$

and

$$\exp(-e^{\phi(x)-t}) = \exp(-e^{\phi(\bar{x})-t} e^{\phi(x)-\phi(\bar{x})}) = z^{\exp(\phi(x)-\phi(\bar{x}))}.$$

# Proof of the distribution of the $\arg\max$

Inserting, we conclude

$$
\begin{aligned}
\mathrm{P}[\arg\max_x \gamma(x) = \bar{x}] &= \mathrm{P}[\gamma(\bar{x}) \geq \max_{x \neq \bar{x}} \gamma(x)] \\
&= \int_{-\infty}^{\infty} g(t; \phi(\bar{x})) \prod_{x \neq \bar{x}} G(t; \phi(x)) \, \mathrm{d}t \\
&= \int_{-\infty}^{\infty} e^{\phi(\bar{x}) - t} \exp(-e^{\phi(\bar{x}) - t}) \prod_{x \neq \bar{x}} \exp(-e^{\phi(x) - t}) \, \mathrm{d}t \\
&= \int_0^1 \prod_{x \neq \bar{x}} z^{\exp(\phi(x) - \phi(\bar{x}))} \, \mathrm{d}z = \int_0^1 z^{\sum_{x \neq \bar{x}} e^{\phi(x) - \phi(\bar{x})}} \, \mathrm{d}z \\
&= \frac{1}{1 + \sum_{x \neq \bar{x}} e^{\phi(x) - \phi(\bar{x})}} = \frac{\exp(\phi(\bar{x}))}{\sum_x \exp(\phi(x))}.
\end{aligned}
$$

# Summary

- Take a Gibbs distribution $\propto e^{\phi(x)}$, parametrized by potential function $\phi$
- Perturb the distribution by adding noise $\epsilon_x \sim \mathrm{Gumbel}(0)$ to every $\phi(x)$
- Optimize the new distribution $\propto e^{\phi(x)+\epsilon_x}$
- The $\arg\max$ will be distributed according to $\propto e^{\phi(x)}$

# Reduced-order Gumbel Perturbation

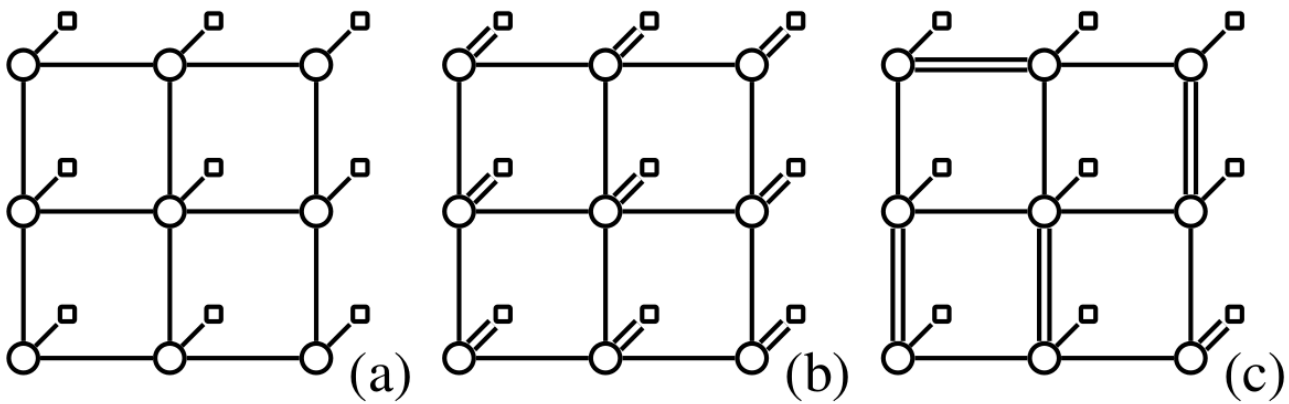
- However, perturbation of the fully-expanded potential table is not practically applicable for the main reasons:
  i) Too many Gumbel samples needed
  ii) No structure of the energy function is exploited for optimization
- Sampling by Optimization is only practical when an efficient optimization algorithm exists

# Example: Ising Model - Illustration

The Ising model uses a potential function

$$-\phi_\theta(x) = E(x,\theta) = \sum_{i=1}^{d} \lambda_i x_i + \sum_{i \sim j} \mu_{ij} x_i x_j,$$

which is a sum of unary and binary potential functions linear in its parameters $\theta = (\lambda, \mu)$.




(a) Original model (b) Order-1 perturbation (c) Order-2 perturbation

# Example: Markov Random Fields and Ising Model

- For the binary case, one needs to make sure that the energy function remains submodular with high probability, i.e. the $\mu_{ij}$ should stay non-negative.
  - Strong links are preferred for perturbation.

- The perturbed energy can then be efficiently minimized by means of powerful optimization algorithms (commonly known as "graphcuts").

Bounds on the approximation
(Hazan, Jaakkola, ICML 2012)

# Theoretical results on low order perturbations

Bounding the error of using lower order perturbation samples for:

- Estimating the Maximum-Likelihood parameter $\theta = \arg\max_\theta p\left(\text{Data}|\theta\right)$
- Estimating the partition function $Z = \sum_{x \in X} e^{\phi(x)}$

Error on the ML estimate will depend on the error on the partition function

# Estimating the partition function

$$\log Z = \mathbb{E}\left[\max_{x \in X}\{\phi(x) + \epsilon_x\}\right] + \text{known constant}$$

Proof:

$$\begin{aligned} P\left(\max_{x \in X}\{\phi(x) + \epsilon_x\} < t\right) &= \Pi_{x \in X} P\left(\epsilon_x < t - \phi(x)\right) \\ &= \Pi_{x \in X} \exp\left(-e^{-t+\phi(x)}\right) \\ &= \exp\left(-\sum_{x \in X} e^{-t+\phi(x)}\right) \\ &= \exp\left(-e^{-t} \cdot e^{-\log Z}\right) \end{aligned}$$

Hence, $\max_{x \in X}\{\phi(x) + \epsilon_x\} \sim \text{Gumbel}(\log Z)$ and its expectation is (a known constant shy of) $\log Z$

# General order perturbations

Let $x = (x_1, \ldots, x_n) \in X = X_1 \times \cdots \times X_n$ and let a family of subsets $\alpha \in \mathcal{A}$ such that $\cup_{\alpha \in \mathcal{A}} = \{1, \ldots, n\}$
General order perturbations are defined by:

$$\max_x \left\{ \phi(x) + \sum_{\alpha \in \mathcal{A}} \epsilon_\alpha(x_\alpha) \right\}$$

For instance in the Ising model

$$-\phi_\theta(x) = \sum_{i=1}^{d} V_i(x_i) + \sum_{i \sim j} V_{ij}(x_i x_j).$$

- Unary perturbations: $\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$ with $\alpha_i = \{i\}$
- Binary perturbations: unary sets $\cup$ $\alpha_{i,j} = \{i, j\}$, for $(i, j)$ neighbors
- Perturbation of the full potential table: only one set $\alpha = \{1, \ldots, n\}$

# Results for partition function estimation

Estimating the partition function with samples generated from a general order perturbation of family $\mathcal{A}$ yields an upper bound:

$$\log Z \leq \mathbb{E}\left[\max_x \{\phi(x) + \sum_{\alpha \in \mathcal{A}} \epsilon_\alpha(x_\alpha)\}\right]$$

Result derives from this identity:

$$\log Z = \mathbb{E}_{\epsilon_1} \max_{x_1} \ldots \mathbb{E}_{\epsilon_n} \max_{x_n} \{\phi(x) + \sum_{i=1}^{N} \epsilon_i(x_i)\}$$

Unary: swap expectations and maximizations to get the upper bound
Disjoint subsets $\alpha$: same
General: need a transformation of the potentials

# Proof of the decomposition of $Z$

$$\log Z = \log \sum_{x_1} \cdots \sum_{x_n} e^{\phi(x)}$$

$$= \log \sum_{x_1} e^{\phi_1(x_1)} \qquad \text{(with } \phi_1(.) = \log \textstyle\sum_{x_2} \cdots \sum_{x_n} e^{\phi(.,x_2,\ldots,x_n)}\text{)}$$

$$= \mathbb{E}_{\epsilon_1} \max_{x_1} \{\log \sum_{x_2} \cdots \sum_{x_n} e^{\phi(x_1,x_2,\ldots,x_n)} + \epsilon_1(x_1)\}$$

$$= \mathbb{E}_{\epsilon_1} \max_{x_1} \ldots \mathbb{E}_{\epsilon_n} \max_{x_n} \{\phi(x) + \sum_{i=1}^{N} \epsilon_i(x_i)\} \quad \text{(follows by induction)}$$

# Parameter estimation (1/2)

Let $\phi_\theta(x) = \theta^T \psi(x)$ a linear parametric potential over features $\psi$ of $x$, and $D = \{x^1 \dots x^K\}$ a dataset. The ML parameter is:

$$\begin{aligned}\theta^* &= \arg\max_\theta \frac{1}{K} \log \Pi_{i=1}^K e^{\phi_\theta(x^i)} * Z^{-1} \\ &= \arg\max_\theta \frac{1}{K} \sum_{i=1}^K \theta^T \psi(x) - \log Z\end{aligned}$$

By maximizing the surrogate function:

$$J(\theta) = \arg\max_\theta \frac{1}{K} \sum_{i=1}^K \theta^T \psi(x) - \mathbb{E}[\max_x \{\phi_\theta(x) + \sum_{\alpha \in \mathcal{A}} \epsilon_\alpha(x_\alpha)\}]$$

It immediately follows (from the bounds on the partition function) that $J(\theta)$ is a lower bound of the data log-likelihood

# Parameter estimation (2/2)

Recall:

$$J(\theta) = \arg\max_{\theta} \frac{1}{K} \sum_{i=1}^{K} \theta^T \psi(x) - \mathbb{E}[\max_{x}\{\phi_\theta(x) + \sum_{\alpha \in \mathcal{A}} \epsilon_\alpha(x_\alpha)\}]$$

$J$ is concave in $\theta$ and smooth with moment matching gradient:

$$\frac{\partial}{\partial \theta} J(\theta) = \frac{1}{K} \sum_{i=1}^{K} \psi(x) - \mathbb{E}[\psi(\hat{x})]$$

Where $\hat{x} \sim \arg\max_x \{\theta^T \cdot \psi(x) + \sum_{\alpha \in \mathcal{A}} \epsilon_\alpha(x_\alpha)\}$ are approximate samples of the model

Exact samples from continuous distributions
(Maddison, Tarlow, Minka, NIPS 2014)

# Gumbel-Max Trick

Consider again a distribution with density

$$p(x) \propto \exp(\phi(x))$$

for any $x \in X$. As before, let $\epsilon_x \sim \mathrm{Gumbel}(0)$ be i.i.d. samples and put

$$\gamma(x) = \phi(x) + \epsilon_x \sim \mathrm{Gumbel}(\phi(x)).$$

If $X$ is continuous, then we have for any $B \subseteq X$

$$\max_{x \in B} \gamma(x) \sim \mathrm{Gumbel}\left(\log \int_{x \in B} \exp(\phi(x))\right),$$

$$\arg\max_{x \in B} \gamma(x) \sim \frac{\exp(\phi(x))}{\int_{x \in B} \exp(\phi(x))}.$$

In particular, these quantities are independent random variables!

# Decomposition

Goal: find $\arg\max \gamma(x)$
Suppose we can decompose

$$\phi(x) = \tau(x) + \beta(x)$$

into a tractable part (i.e. allows efficient sampling) $\tau(x)$ and a bounded part $\beta(x)$ which satisfies

$$\beta(x) \leq M(B) \quad \forall x \in B \subseteq X,$$

where $M(B)$ is a tractably computable bound depending on the region $B$.
Recover the Gumbel noise sample $\gamma(x)$ by adding the difference $\beta(x)$ to a tractable sample from $\mathrm{Gumbel}(\tau(x))$. More precisely,

$$\gamma(x) = \epsilon_x + \phi(x) = \epsilon_x + \tau(x) + \beta(x) \sim \mathrm{Gumbel}(\tau(x)) + \beta(x).$$

# Upper and Lower Bounds

Thus, an upper bound $u_B$ on our target value $\max_{x \in B} \gamma(x)$ can be obtained via

$$\max_{x \in B} \gamma(x) = \max_{x \in B} \epsilon_x + \tau(x) + \beta(x) \leq \max_{x \in B} \epsilon_x + \tau(x) + M(B) =: u_B,$$

while a lower bound (over the whole space $X$) is provided by

$$\ell := \gamma(\bar{x}) = \epsilon_{\bar{x}} + \tau(\bar{x}) + \beta(\bar{x}), \forall \bar{x} \in X$$

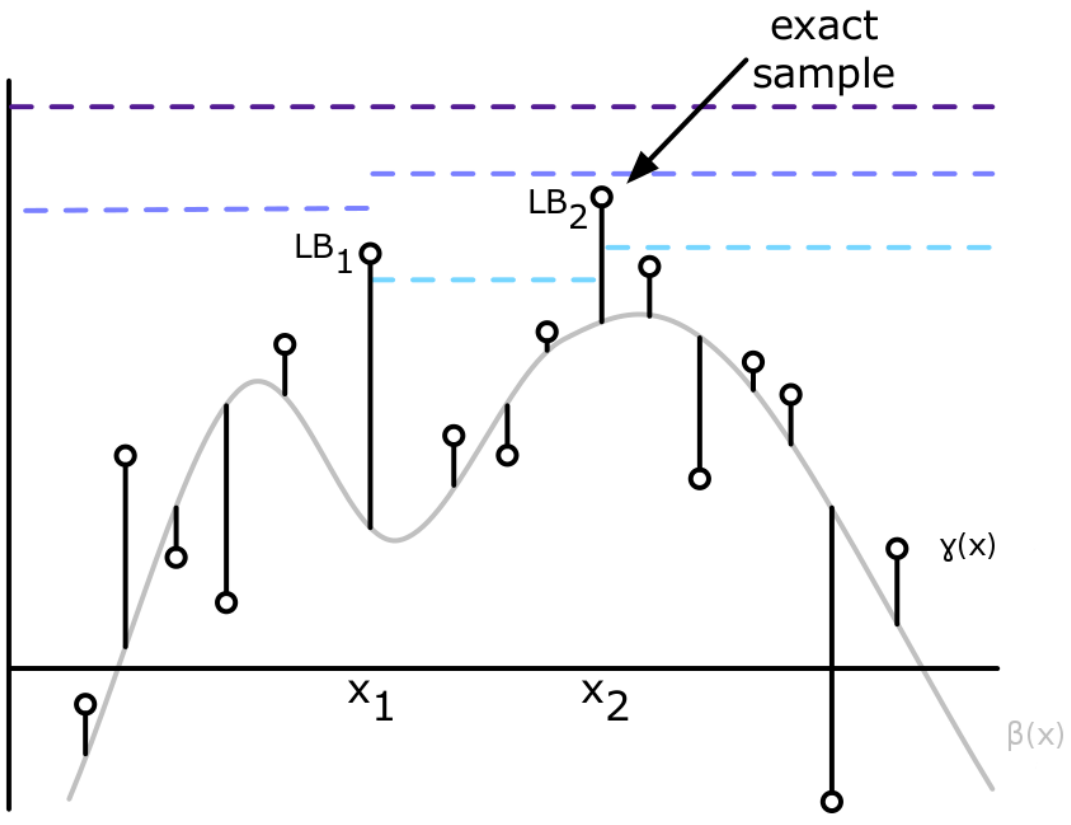
This allows for finding $\arg\max_x \gamma(x)$ by applying a procedure similar to branch-and-bound.

# Sampling Algorithm

Basic idea:

1. Initialize lower and upper bounds for the target quantity.
2. Sample $\epsilon_x + \tau(x)$ on subregions $B$.
3. Partition the space into further subsets and update bounds (such that the gap becomes narrower).
4. Discard subregions where the lower bound violates the upper bound.
5. Proceed recursively until no more subregions are left and return the optimal point.

# Visualization




Algorithm:

1. Initialize lower and upper bounds $\ell$ and $u_X$.
2. If $\ell \leq u_B$ for $B$, sample $\bar{x} = \arg\max_{x \in B} \epsilon_x + \tau(x)$.
3. If $\ell \leq \gamma(\bar{x})$, update $\ell$ and proceed.
4. Partition into subspaces $L, R$ of $B$ and update bounds $u_L, u_R$.
5. Proceed recursively until all regions are searched.

# Summary

- When the bounding function $M(B)$ does not depend on the region, behavior similar to rejection sampling
- Else, experimentally demonstrated to be more efficient than adaptive rejection sampling
- In all cases, it is only reasonable for small dimensions

# Conclusion

We have seen algorithms for:

- Approximate sampling from Gaussian MRF, useful in high dimensions
- Approximate sampling from Gibbs distributions with potential linear in the features (and some guaranties on the sampling)
- Exact sampling from continuous distributions with small dimensions where a bounding function $M(B)$ can be defined

# References


[1] *Gaussian sampling by local perturbations*, Papandreou, Yuille, NIPS 2010

[2] *Perturb-and-MAP Random Fields: Using Discrete Optimization to Learn and Sample from Energy Models*, Papandreou, Yuille, ICCV 2011

[3] *On the Partition Function and Random Maximum A-Posteriori Perturbations*, Hazan, Jaakkola, ICML 2012

[4] $A^*$ *Sampling*, Maddison, Tarlow, Minka, NIPS 2014