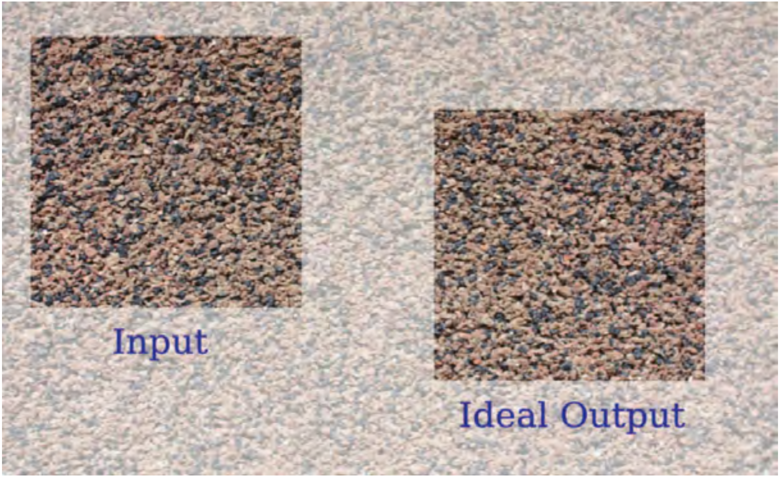
Maximum Entropy Distributions For Image Synthesis Under Statistical Constraints

Agnès Desolneux

CNRS and Ecole Normale Supérieure Paris-Saclay

Journée Statistique et Informatique à Paris Saclay
Vendredi 5 février 2021

The problem of exemplar-based texture synthesis



Sampling under statistical constraints

Given an image $I_0 \in \mathbb{R}^d$, and some statistical features (empirical mean, covariance, etc.) $F(I_0) \in \mathbb{R}^p$, the goal is to sample an image I such that

$$\text{“} \quad F(I) = F(I_0) \quad \text{”}.$$

Sampling under statistical constraints

Given an image $I_0 \in \mathbb{R}^d$, and some statistical features (empirical mean, covariance, etc.) $F(I_0) \in \mathbb{R}^p$, the goal is to sample an image I such that

$$\text{“} \quad F(I) = F(I_0) \quad \text{”}.$$

Stochastic framework : look for a probability measure P on images $X \in \mathbb{R}^d$ such that P is of **maximum entropy and** such that

$$F(X) = F(I_0) \text{ almost surely when } X \sim P \quad \textbf{(Microcanonical model)}$$

$$\text{or} \quad \mathbb{E}_P(F(X)) = F(I_0) \quad \textbf{(Macrocanonical model)},$$

(def by [Bruna-Mallat 2018]).

The notion of **entropy** : when P has a probability density f_P with respect to the Lebesgue measure dx on $\mathbb{R}^d$, then its (differential) entropy is

$$H(P) = -\int_{\mathbb{R}^d} f_P(x) \log f_P(x)\, dx.$$

J. Bruna and S. Mallat. Multiscale Sparse Microcanonical Models. *Mathematical Statistics and Learning*, 2018.

Macrocanonical models : exponential models

For $\theta \in \mathbb{R}^P$, let us define the exponential model P_θ by

$$f_{P_\theta}(x) = \frac{1}{Z(\theta)} e^{-<\theta, F(x) - F(I_0)>},$$

where

$$Z(\theta) = \int_{\mathbb{R}^d} e^{-<\theta, F(y) - F(I_0)>} \, dy.$$

Notice that

$$\nabla \log Z(\theta) = -\mathbb{E}_{P_\theta}(F(X) - F(I_0)),$$

and

$$D^2 \log Z(\theta) = \mathbb{E}_{P_\Theta}\left[(F(X) - F(I_0))(F(X) - F(I_0))^t\right].$$

Macrocanonical models : exponential models

For $\theta \in \mathbb{R}^P$, let us define the exponential model P_θ by

$$f_{P_\theta}(x) = \frac{1}{Z(\theta)} e^{-<\theta, F(x)-F(I_0)>},$$

where

$$Z(\theta) = \int_{\mathbb{R}^d} e^{-<\theta, F(y)-F(I_0)>}\, dy.$$

Notice that

$$\nabla \log Z(\theta) = -\mathbb{E}_{P_\theta}(F(X) - F(I_0)),$$

and

$$D^2 \log Z(\theta) = \mathbb{E}_{P_\Theta}\left[(F(X) - F(I_0))(F(X) - F(I_0))^t\right].$$

Therefore

$$\theta \mapsto \log Z(\theta) \text{ is convex}$$

and its minimum is achieved (if it exists) at a $\theta^\star$ such that $\mathbb{E}_{P_{\theta^\star}}(F(X)) = F(I_0)$.

Assume that we have $\theta^\star$ such that $\mathbb{E}_{P_{\theta^\star}}(F(X)) = F(I_0)$.

Now let P be another distribution that satisfies $\mathbb{E}_P(F) = F(I_0) = \mathbb{E}_{P_{\theta^\star}}(F)$. Then

$$\mathrm{KL}(P \| P_{\theta^\star}) = \int_{\mathbb{R}^d} f_P(x) \log \frac{f_P(x)}{f_{P^\star_\theta}(x)}\, dx = -H(P) + H(P_{\theta^\star}) \geqslant 0.$$

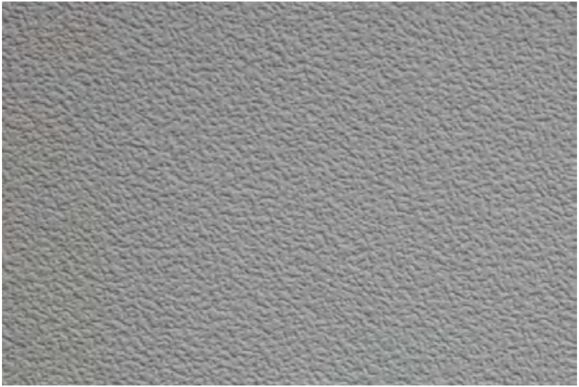
Therefore $P_{\theta^\star}$ is of maximum entropy among all distributions that satisfy $\mathbb{E}_P(F) = F(I_0)$.

But depending on F, it's not always easy to find $\theta^\star$ and to sample from P_θ.

S. C. Zhu, Y. N. Wu, and D. Mumford, Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling, *International Journal of Computer Vision*, 1998.

1. Gaussian Texture Synthesis

First and second-order statistics



First and second-order statistics

Starting from an exemplar texture image $I_0 \in \mathbb{R}^d$, defined on a rectangular domain $\Omega \subset \mathbb{Z}^2$, where $d = |\Omega|$ is the number of pixels, one can consider its average gray-level m_0 and its auto-correlation matrix C_0 given by

$$m_0 = \frac{1}{d} \sum_{k \in \Omega} I_0(k),$$

$$\text{and} \quad C_0(k,l) = \frac{1}{d} \sum_{n \in \Omega} (I_0(k+n) - m_0)(I_0(l+n) - m_0),$$

where I_0 is extended to $\mathbb{Z}^2$ by periodicity.

Then the macrocanonical distribution under the constraints $\mathbb{E}(X) = (m_0 \ldots, m_0)$ and $\mathrm{Cov}(X) = C_0$ is the Gaussian distribution of mean $\mathbf{m_0} = (m_0, \ldots, m_0) \in \mathbb{R}^d$ and covariance matrix C_0.

How to sample from this Gaussian distribution ?

Simply take a discrete white noise W on Ω, i.e. the $W(k)$ are i.i.d. $\mathcal{N}(0,1)$. Then define

$$X = \mathbf{m_0} + \frac{1}{\sqrt{d}}(I_0 - \mathbf{m_0}) * W,$$

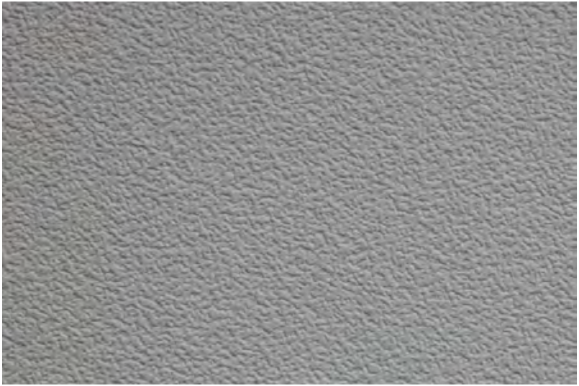
where $*$ is the (discrete periodic) convolution :

$$(I * W)(k) = \sum_{n\in\Omega} I(k-n)W(n).$$

$\Longrightarrow$ We have that $X \sim \mathcal{N}(\mathbf{m_0}, C_0)$.

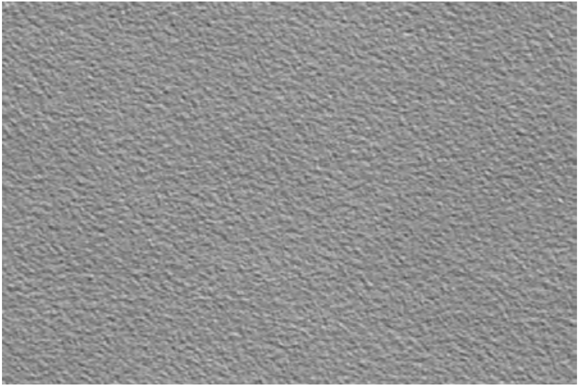
B. Galerne, Y. Gousseau, J.-M. Morel, Random Phase Textures : Theory and Synthesis, *IEEE Transactions on Image Processing*, 2011.

First and second-order statistics : example



Original image I_0

First and second-order statistics : example

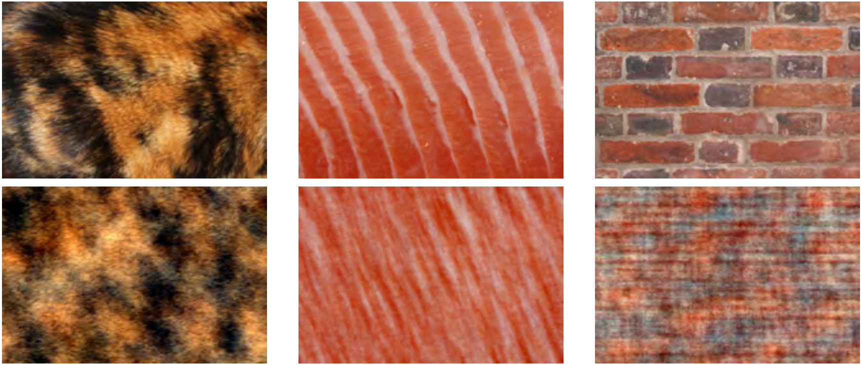


Sample Image $I \sim \mathcal{N}(\mathbf{m_0}, C_0)$

Some examples of « microtextures »

Test the algorithm online at `www.ipol.im/pub/art/2011/ggm_rpn/`.

Some more examples



2. Texture synthesis from CNN features

The framework

Fix $x_0 \in \mathbb{R}^d$ (exemplar texture), μ a reference measure on $\mathbb{R}^d$ and $F: \mathbb{R}^d \to \mathbb{R}^p$ (constraints) with $p \ll d$.

- $\mathbb{R}^d$: image space,
- $\mathbb{R}^p$: parameter space.

The notion of entropy is replaced by the KL divergence from a reference probability measure μ, i.e. consider

$$KL(P\|\mu) = \int_{\mathbb{R}^d} \frac{dP}{d\mu}(x) \log \frac{dP}{d\mu}(x)\; d\mu(x), \text{ when } P \ll \mu \text{ and } +\infty \text{ otherwise.}$$

Consider exponential models : for $x \in \mathbb{R}^d$, define

$$\frac{\mathrm{d}\Pi_\theta}{\mathrm{d}\mu}(x) = \frac{\exp[-\langle \theta, F(x) - F(x_0)\rangle]}{\int_{\mathbb{R}^d} \exp[-\langle \theta, F(y) - F(x_0)\rangle]\, dy}.$$

Question : can we get practical conditions for the existence of a macrocanonical model $\Pi^\star = \Pi_{\theta^\star}$?

Turning to the *finite dimensional* and *convex* dual problem, we get :

Proposition : If there exists $\alpha > 0$ such that

- for all $\eta > 0$, $\int_{\mathbb{R}^d} \exp[\eta \|x\|^\alpha] \mathrm{d}\mu(x) < +\infty$;
- $\sup_{x \in \mathbb{R}^d} \{F(x)(1 + \|x\|^\alpha)^{-1}\} < +\infty$;
- for all $\theta \in \mathbb{S}^{p-1}$, $\mu(\{x \in \mathbb{R}^d ; \langle \theta, F(x) - F(x_0) \rangle < 0\}) > 0$,

then there exists $\theta^\star \in \mathbb{R}^p$ such that $\Pi_{\theta^\star}$ is a macrocanonical model associated with F and x_0.

1. How to find the optimal $\theta^\star$?
2. How to sample from the model $\frac{\mathrm{d}\Pi_\theta}{\mathrm{d}\mu}(x) \propto \exp[-\langle \theta, F(x) - F(x_0) \rangle]$?

V. De Bortoli, A. Desolneux, A. Durmus, B. Galerne and A. Leclaire. Maximum entropy methods for texture synthesis : theory and practice. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 2020.

Finding the optimal parameters . . .

The optimal parameters $\theta^\star$ minimize the *log-partition function*

$$L(\theta) = \log \left[\int_{\mathbb{R}^d} \exp(-\langle \theta, F(x) - F(x_0) \rangle) \mathrm{d}\mu(x) \right].$$

Properties of the log-partition function :

- $\nabla_\theta L(\theta) = -\mathbb{E}_{\Pi_\theta}(F - F(x_0))$,
- $\nabla_\theta^2 L(\theta) = \mathrm{Cov}_{\Pi_\theta}(F - F(x_0)) \Rightarrow$ convexity

Gradient descent :

$$\theta_{n+1} = \mathrm{Proj}_K[\theta_n + \delta_{n+1} \mathbb{E}_{\Pi_{\theta_n}}(F - F(x_0))],$$

where K is some compact convex set which contains $\theta^\star$.

But need to compute $\mathbb{E}_{\Pi_\theta}(F) \Rightarrow$ *Monte Carlo* approximation of $\mathbb{E}_{\Pi_\theta}(F)$
$\Rightarrow$ how to sample from Π_θ ?

Sampling from Π_θ ...

Usually it is not possible to sample from $\Pi(\mathrm{d}x) \propto \exp[-U(x)]\mathrm{d}x$, **but**, approximate sampling is available (under some conditions) via **Langevin dynamic**

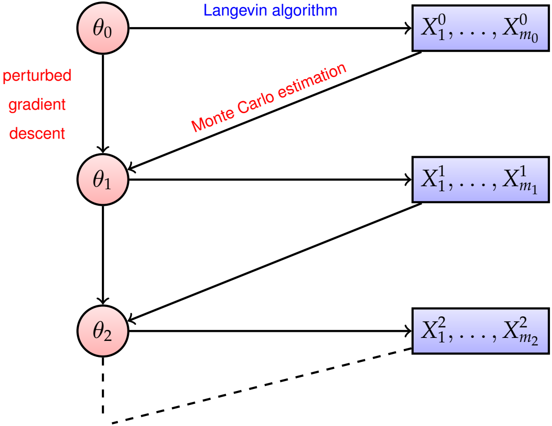
$$X_{n+1} = X_n - \gamma_{n+1}\nabla U(X_n) + \sqrt{2\gamma_{n+1}}Z_{n+1},$$

where $Z_{n+1} \sim \mathcal{N}(0, \mathrm{Id})$, i.i.d., and $\gamma_n > 0$.

$\rightarrow$ **sample** $\approx \Pi(\mathrm{d}x) \propto \exp[-U(x)]\mathrm{d}x$.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 2017.

Combining dynamics



– parameter sequence $\in \mathbb{R}^p$ (optimization)

– image sequence $\in \mathbb{R}^d$ (sampling)

Let $U_\theta(x) = \langle \theta, F(x) - F(x_0) \rangle + r(x)$ (assuming that $\frac{d\mu}{d\text{Leb}}(x) \propto \exp(-r(x))$).

Finding optimal parameters

$\theta^\star$ is the minimum of the log-partition function which is a *convex problem*.

Gradient descent dynamics

$$\theta_{n+1} = \theta_n + \delta_{n+1} \mathbb{E}_{\Pi_{\theta_n}}(F - F(x_0))$$

Sampling from a Gibbs measure

The potential $x \mapsto U_\theta(x)$ is usually *non-convex* but has *curvature at infinity*.

Langevin dynamics

$$X_{n+1} = X_n - \gamma_{n+1} \nabla U_\theta(X_n) + \sqrt{2\gamma_{n+1}} Z_{n+1}$$

$\Rightarrow$ **Combining dynamics**

$$X_{k+1}^n = X_k^n - \gamma_n \nabla U_{\theta_n}(X_k^n) + \sqrt{2\gamma_n} Z_{k+1}^n, \text{ with } X_0^n = X_{m_{n-1}}^{n-1},$$

$$\theta_{n+1} = \text{Proj}_K[\theta_n + \delta_{n+1} m_n^{-1} \sum_{k=1}^{m_n} \{F(X_k^n) - F(x_0)\}],$$

where $Z_k^n \sim \mathcal{N}(0, \text{Id})$, i.i.d.

Main result : De Bortoli, Durmus, Pereyra, Fernandez Vidal (2018).

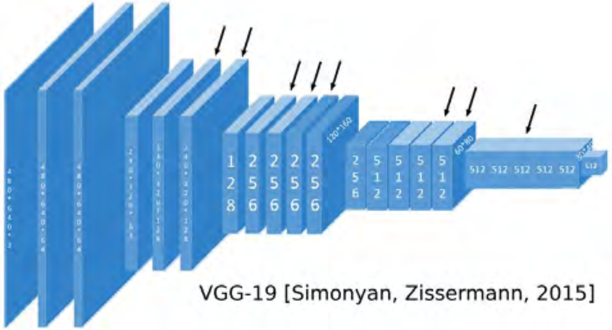
Theorem :[Convergence of the parameters]
If F is smooth and $\|F\| \leqslant V$ with $V : \mathbb{R}^d \to [1, +\infty)$ a Lyapunov function for the Langevin dynamics, and

$$\sum_{k=1}^{+\infty} \delta_k = +\infty, \quad \sum_{k=1}^{+\infty} \frac{\delta_k}{m_k \gamma_k} < +\infty, \quad \sum_{k=1}^{+\infty} \delta_k \gamma_k^{1/2} < +\infty.$$

then $(\theta_n)_{n\in\mathbb{N}}$ converges almost surely and in L^1 to the optimal parameters.

Remark : F does **not** need to be convex (neural network features ✓).

j = family of layers
c_ℓ = channels of layer ℓ
$\mathcal{G}_{\ell,c}$ = CNN feature at layer ℓ and channel c
$n_{\ell,c}$ = number of pixels at layer ℓ and channel c

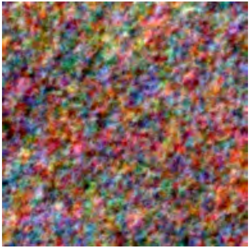


Structure of the neural network VGG-19

Choice of features : mean of each *channel* for selected *layers*, $p \approx 10^3$, i.e. $F(x) = (\sum_{i=1}^{n_{\ell,c}} \mathcal{G}_{\ell,c}(x)_i / n_{\ell,c})_{\ell \in \mathbf{j}, c \in c_\ell}$.

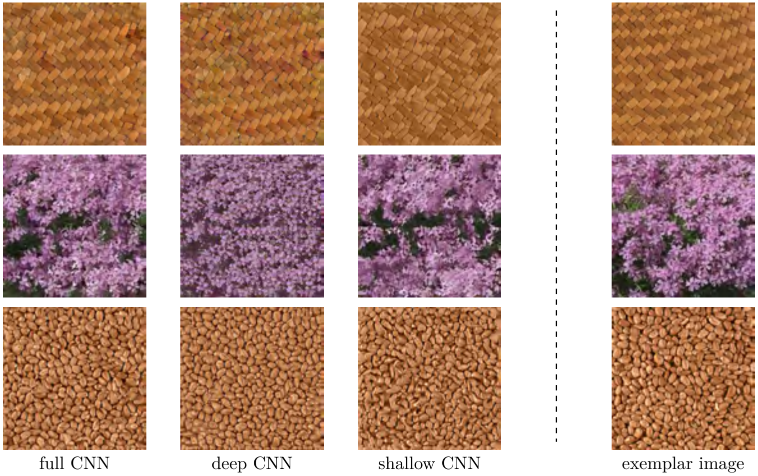
Examples


Input (x_0)


Initialization (Gaussian)


After 10000 iterations

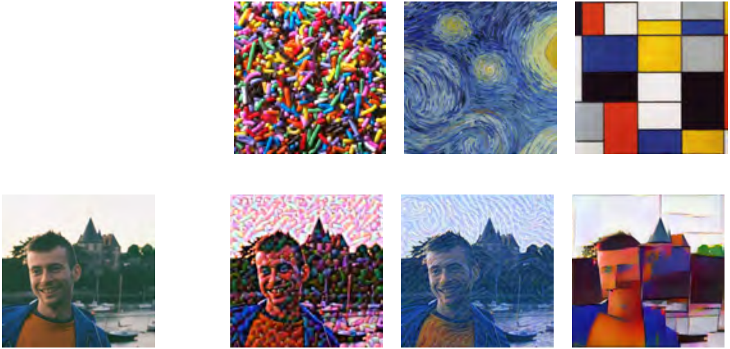
Examples



Examples

Lu-Zhu-Wu | Gatys | ours | exemplar image x_0

Extension : style transfer



exemplar image x_0 (a) (b) (c)

Conclusion

Sampling from maximum entropy models under statistical constraints :

- allows to understand the constraints (features, descriptors, SIFT for instance),
- raises many mathematical questions (existence, sampling algorithms, etc.)
- provides generative models of images (used then as priors ?)