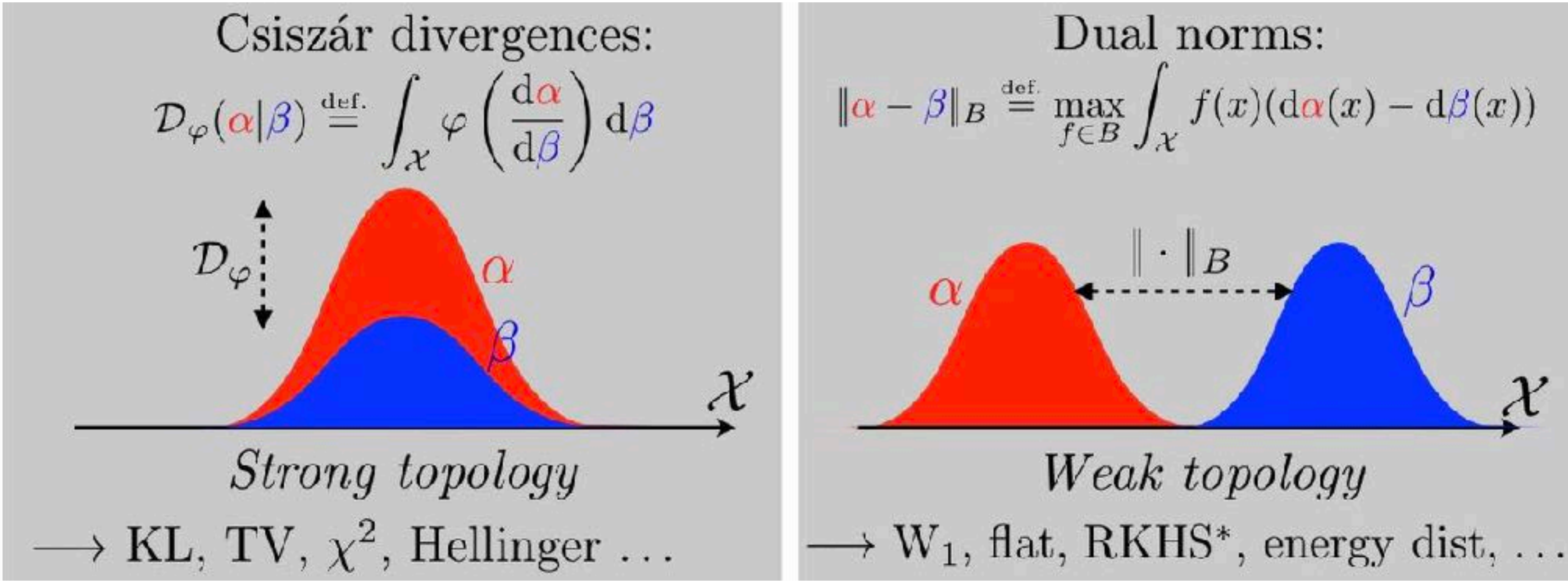
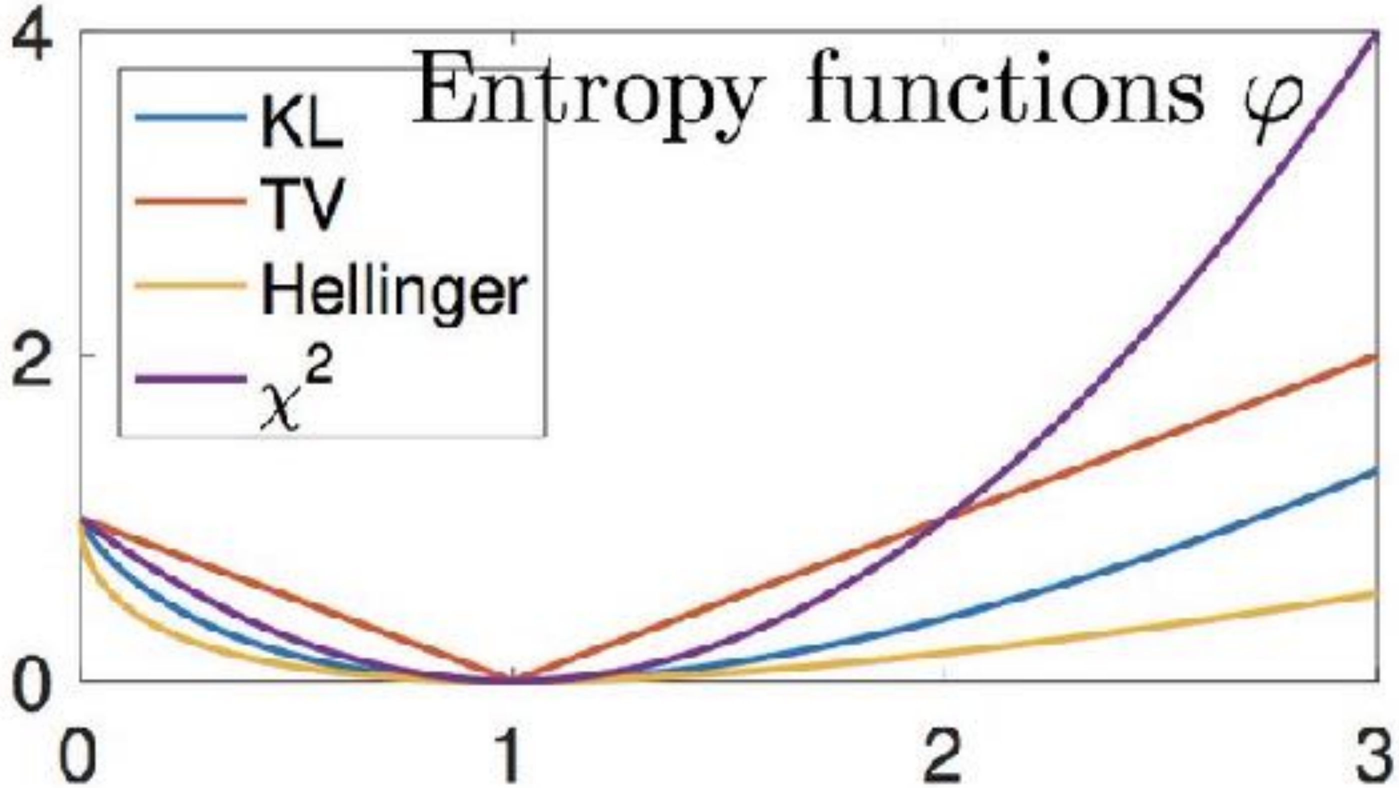
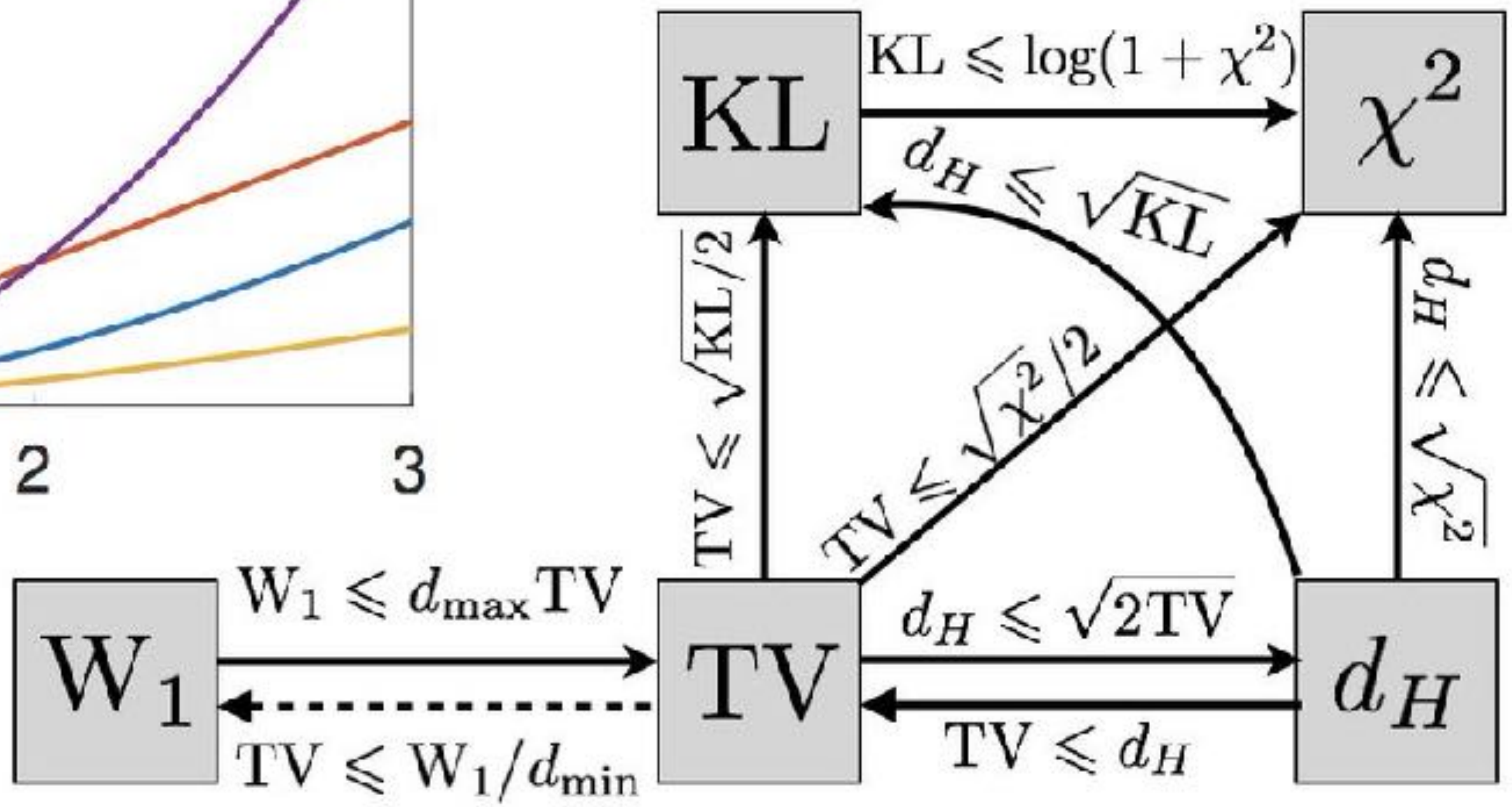
Measures and Probability

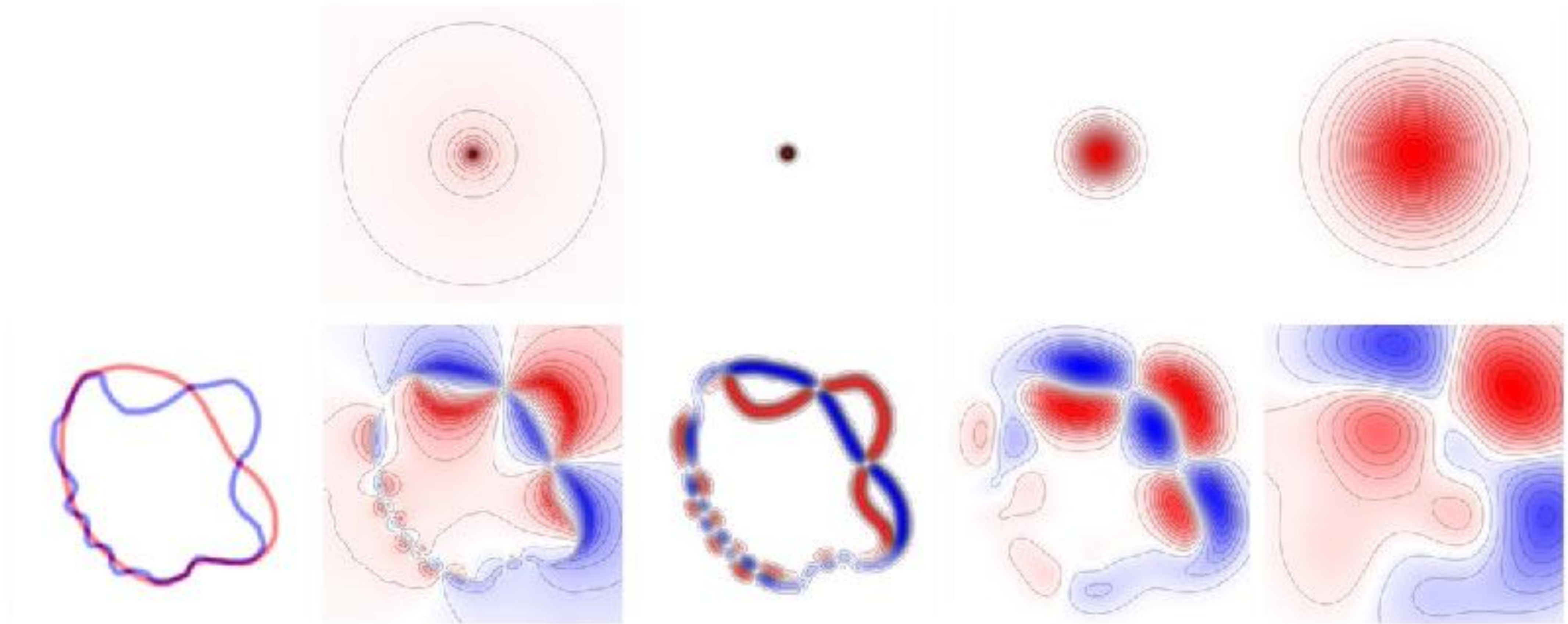

Csiszár divergences:
$\mathcal{D}_\varphi(\alpha|\beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta$
$\mathcal{D}_\varphi$
α
β
$\mathcal{X}$
Strong topology
$\longrightarrow$ KL, TV, χ^2, Hellinger ...
Dual norms:
$\|\alpha - \beta\|_B \stackrel{\text{def.}}{=} \max_{f \in B} \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x))$
$\|\cdot\|_B$
α
β
$\mathcal{X}$
Weak topology
$\longrightarrow$ W_1, flat, RKHS*, energy dist, ...

$$\mathcal{D}_\varphi(\alpha|\beta) \stackrel{\text{def.}}{=} \int_\mathcal{X} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right)\mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X})$$

$$\varphi'_\infty = \lim_{x\uparrow+\infty} \varphi(x)/x \in \mathbb{R}\cup\{\infty\}$$

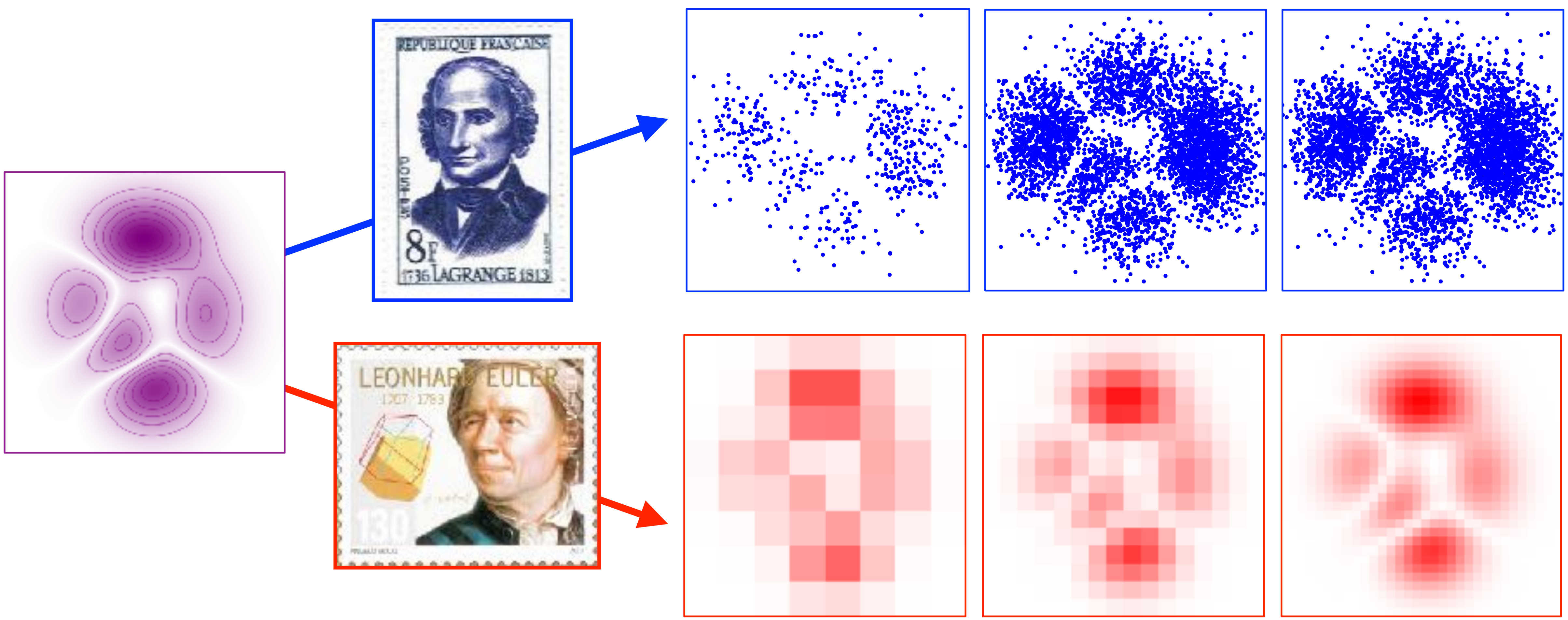






(α, β) ED $(G, .005)$ $(G, .02)$ $(G, .05)$

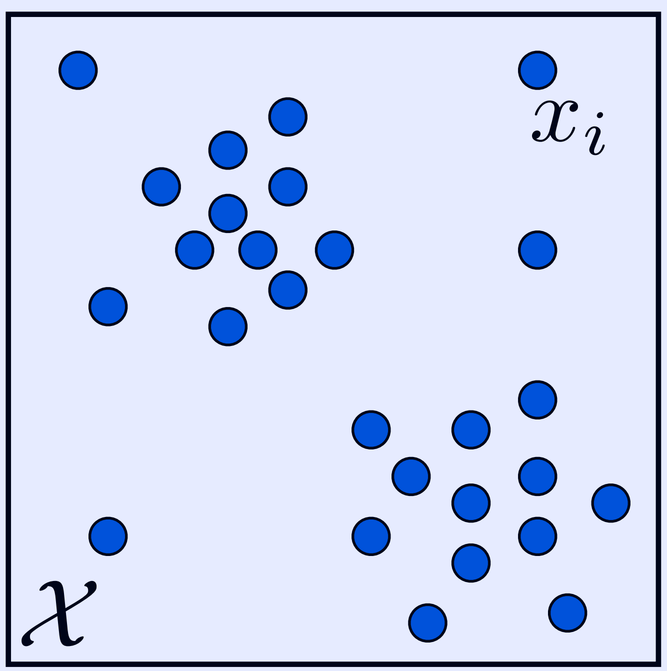
Figure 8.4: Top row: display of ψ such that $\|\alpha - \beta\|_k = \|\psi \star (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{\varphi}(\omega)}$ where $k^*(x, x') = \varphi(x - x')$. Bottom row: display of $\psi \star (\alpha - \beta)$. (G,$\sigma$) stands for Gaussian kernel of variance σ^2 and ED for Energy Distance kernel (in which case $\psi(x) = 1/\sqrt{\|x\|}$).


REPUBLIQUE FRANÇAISE
8F
1736 LAGRANGE 1813
EULER
1707 1783
130

Discrete measure: $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad x_i \in \mathcal{X}, \quad \sum_i \mathbf{a}_i = 1$

Lagrangian (point clouds)

Constant weights $\mathbf{a}_i = \frac{1}{n}$

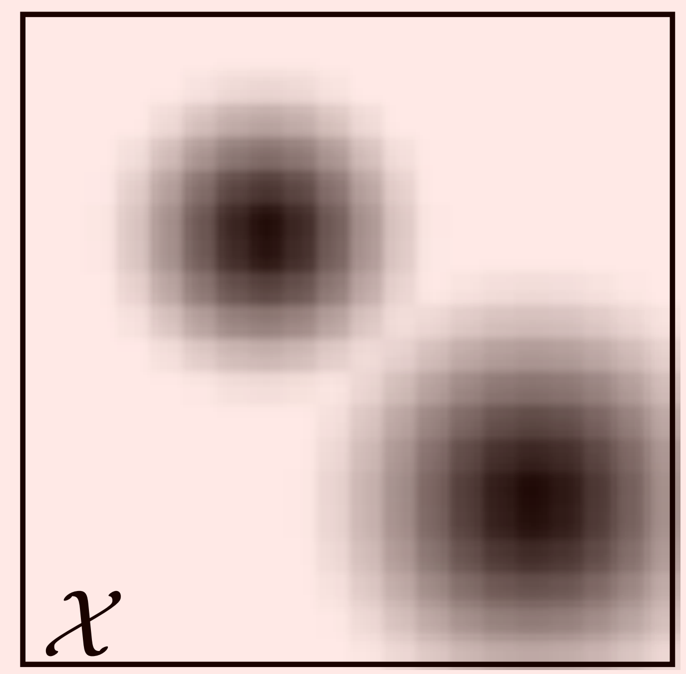


Quotient space:

$\mathcal{X}^n / \mathrm{Perm}(n)$

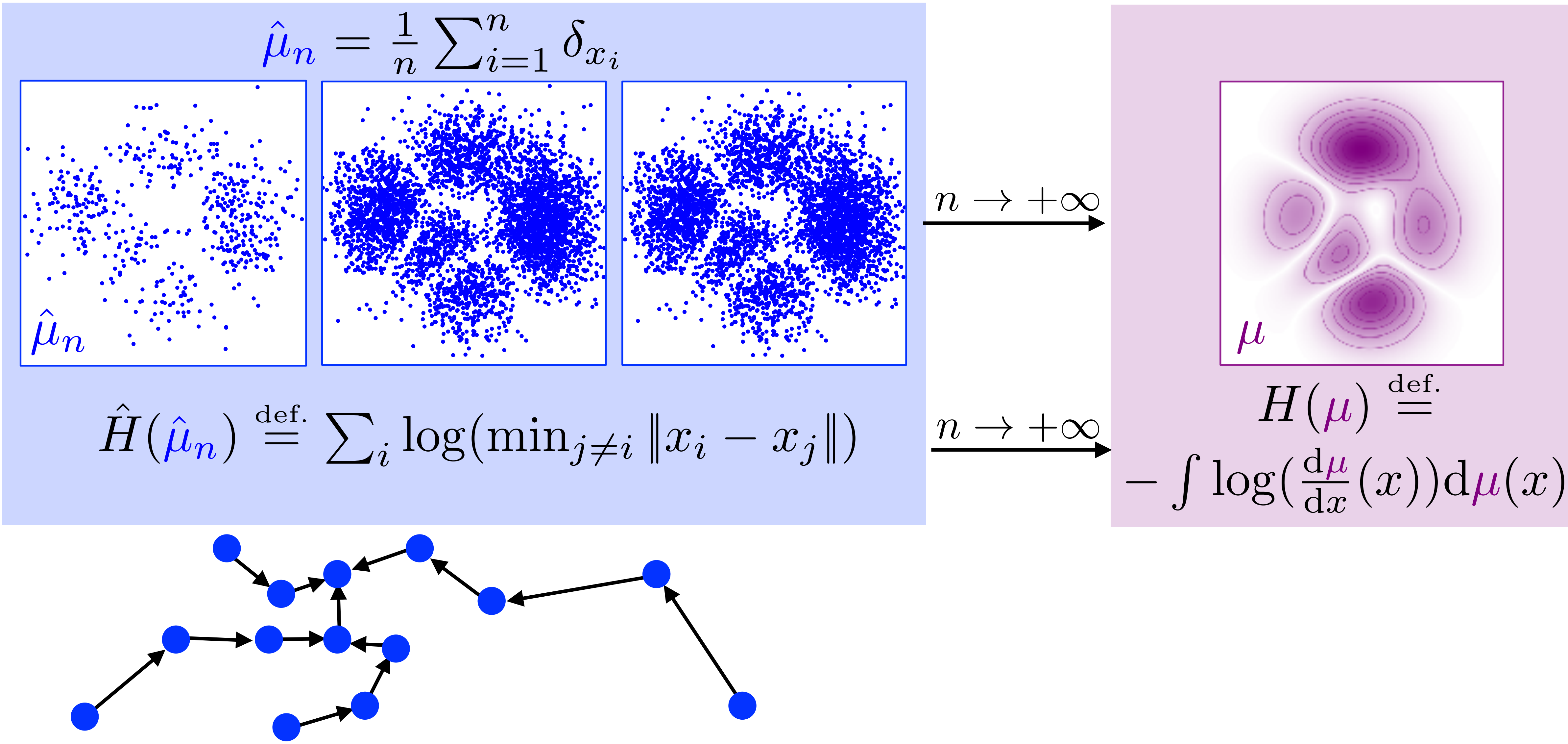
Eulerian (histograms)

Fixed positions x_i (e.g. grid)



Convex polytope (simplex):

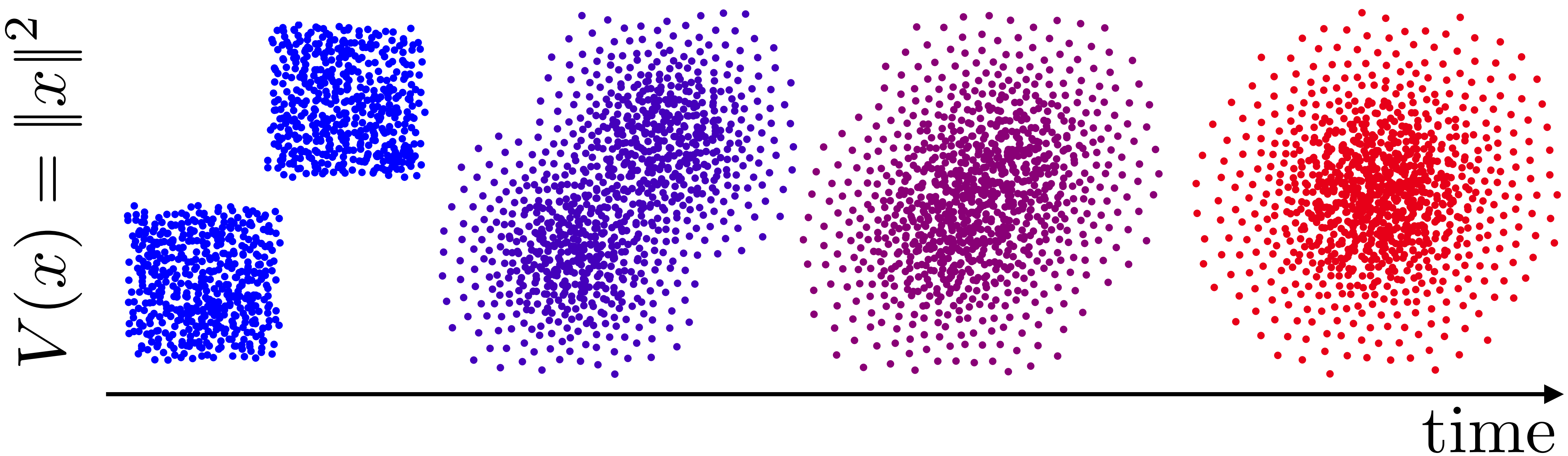
$\{(\mathbf{a}_i)_i \geqslant 0 \; ; \; \sum_i \mathbf{a}_i = 1\}$



$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$
$\hat{\mu}_n$
$n \to +\infty$
μ
$\hat{H}(\hat{\mu}_n) \stackrel{\text{def.}}{=} \sum_i \log(\min_{j \neq i} \|x_i - x_j\|)$
$n \to +\infty$
$H(\mu) \stackrel{\text{def.}}{=} -\int \log(\frac{\mathrm{d}\mu}{\mathrm{d}x}(x))\mathrm{d}\mu(x)$

$$\min_{\rho} E(\rho) \stackrel{\text{def.}}{=} \int V(x)\rho(x)\mathrm{d}x + \int \rho(x)\log(\rho(x))\mathrm{d}x$$

Wasserstein flow of E: $\dfrac{\mathrm{d}\rho_t}{\mathrm{d}t} = \Delta\rho_t + \nabla(V\rho_t)$

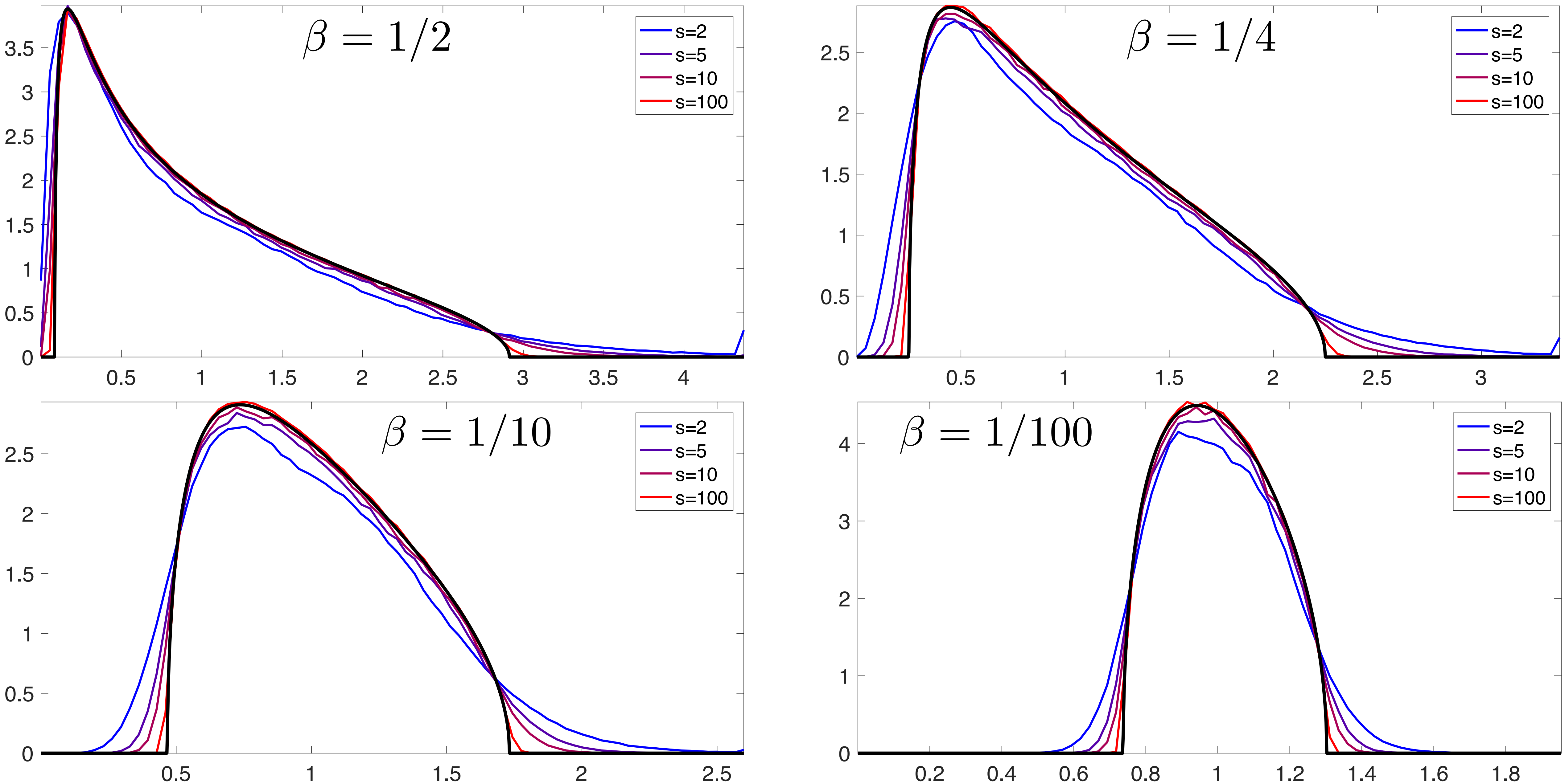


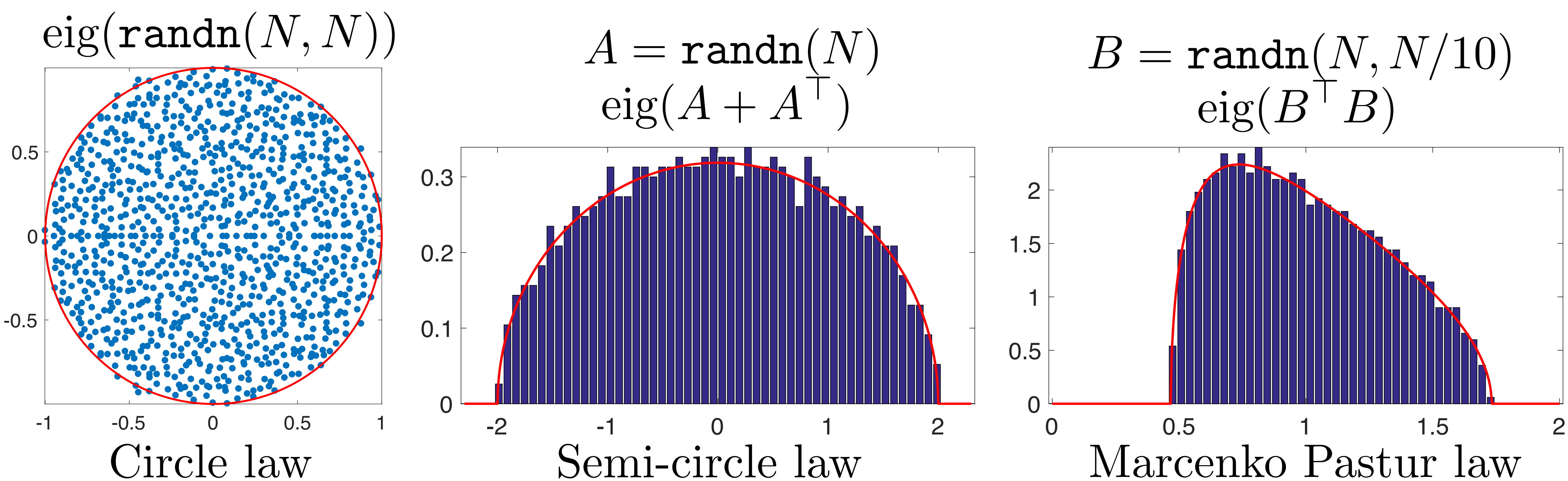
$$B \sim \frac{1}{\sqrt{P}}\texttt{randn}(P,s) \qquad \mathbb{P}(\mathrm{eig}(B^\top B) \in [u,v]) \xrightarrow[s \to +\infty]{\beta \overset{\text{def.}}{=} \frac{s}{P}} \int_u^v f_\beta(\lambda)\mathrm{d}\lambda$$

$$f_\beta(\lambda) \overset{\text{def.}}{=} \frac{1}{2\pi\beta\lambda}\sqrt{(\lambda-\lambda_-)(\lambda_+-\lambda)}1_{[\lambda_-,\lambda_+]}(\lambda)$$

[Marcenko-Pastur]

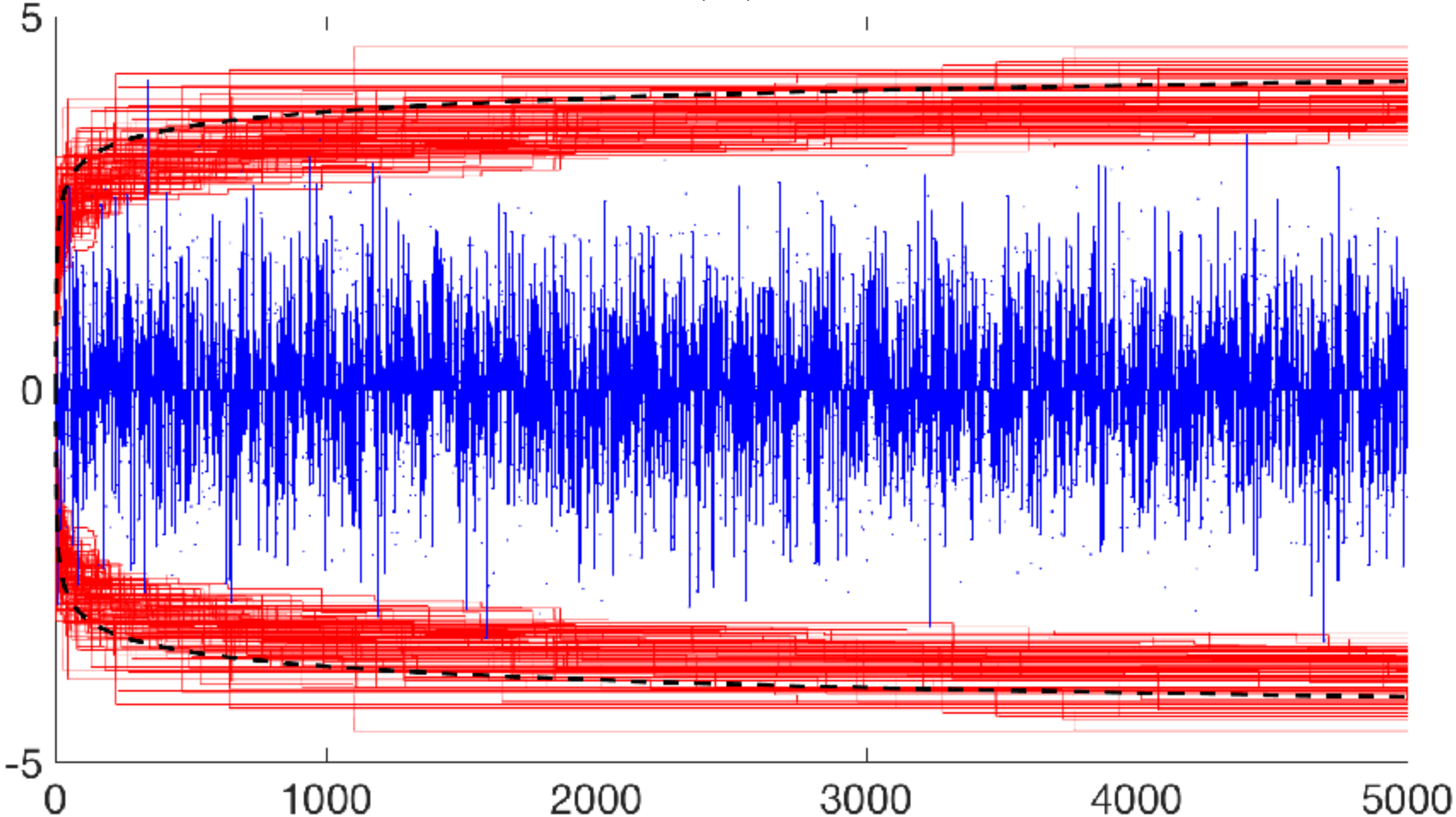
$$\lambda_\pm \overset{\text{def.}}{=} (1 \pm \sqrt{\beta})^2$$



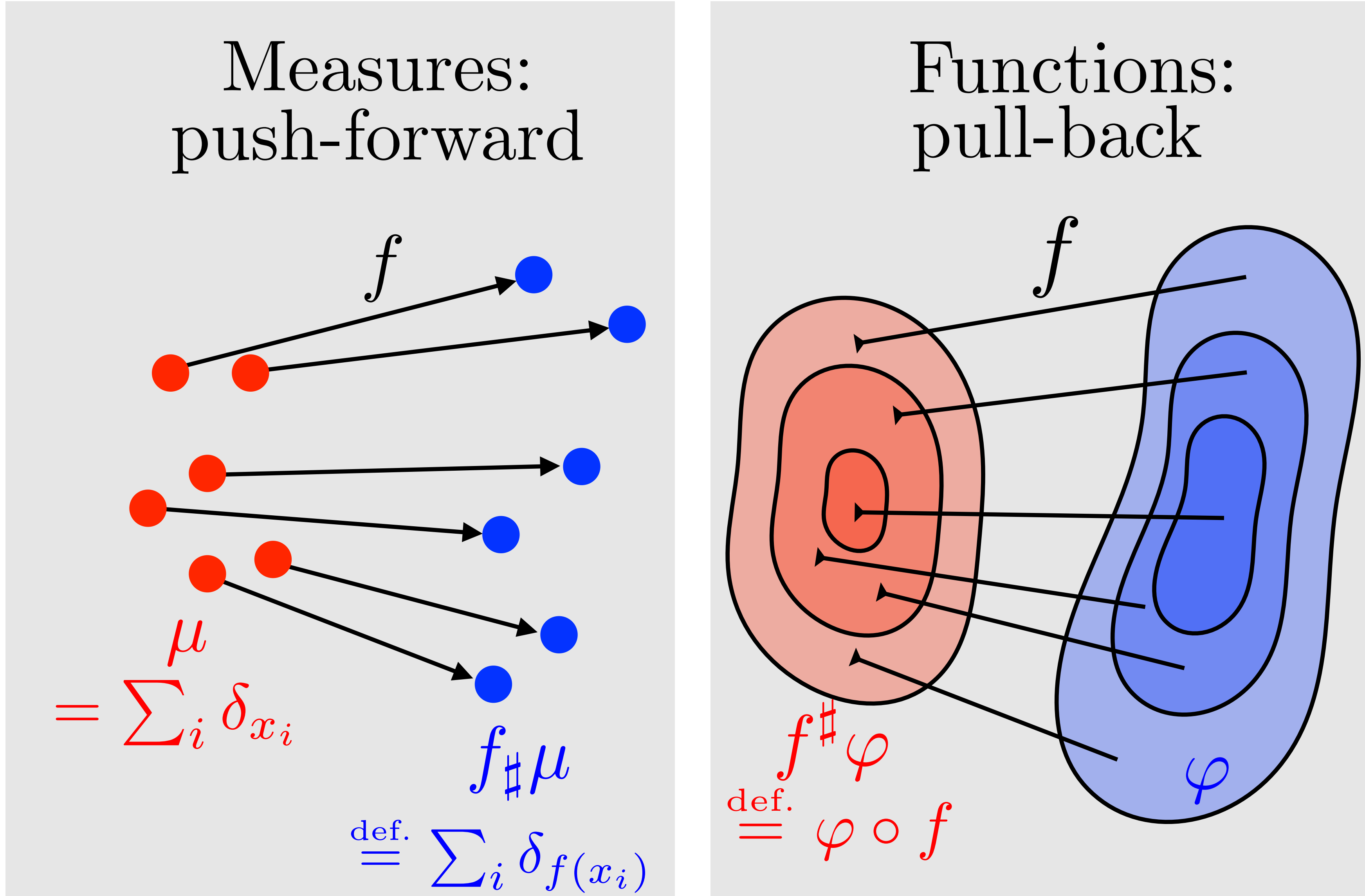


eig(randn(N, N))
0.5
0
-0.5
-1
-0.5
0
0.5
1
Circle law
A = randn(N)
eig(A + A⊤)
0.3
0.2
0.1
0
-2
-1
0
1
2
Semi-circle law
B = randn(N, N/10)
eig(B⊤B)
2
1.5
1
0.5
0
0
0.5
1
1.5
2
Marcenko Pastur law

$$X_i \sim \mathcal{N}(0,1) \qquad \max_{1\leqslant i\leqslant n} |X_i| \sim \sqrt{2\log(n)}$$



$f : \mathcal{X} \to \mathcal{Y}$

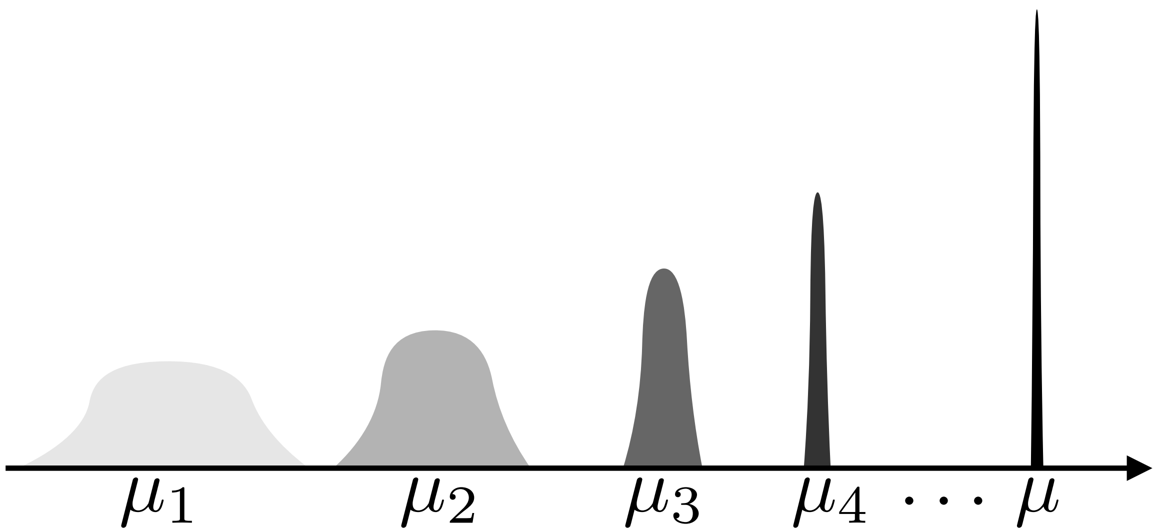


$f_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ $\quad$ $f^\sharp : \mathcal{C}(\mathcal{Y}) \to \mathcal{C}(\mathcal{X})$

Remark: $f^\sharp$ and $f_\sharp$ are adjoints

$$\int_{\mathcal{Y}} \varphi \, \mathrm{d}(f_\sharp \mu) = \int_{\mathcal{X}} (f^\sharp \varphi) \, \mathrm{d}\mu$$

Random vectors		Radon measures
$\mathbb{P}(X \in A)$	$=$	$\int_A \mathrm{d}\mu(x)$
Weak* convergence:		Convergence in law:
$\forall$ set A $\mathbb{P}(X_n \in A) \xrightarrow{n\to+\infty} \mathbb{P}(X \in A)$		$\forall$ continuous function f $\int f \mathrm{d}\mu_n \xrightarrow{n\to+\infty} \int f \mathrm{d}\mu$



In mean

$$\lim_{n\to+\infty} \mathbb{E}(|X_n - X|^p) = 0$$

Almost sure

$$\mathbb{P}(\lim_{n\to+\infty} X_n = X) = 1$$

$\Downarrow$

In probability

$$\forall\, \varepsilon > 0,\ \mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n\to+\infty} 0$$

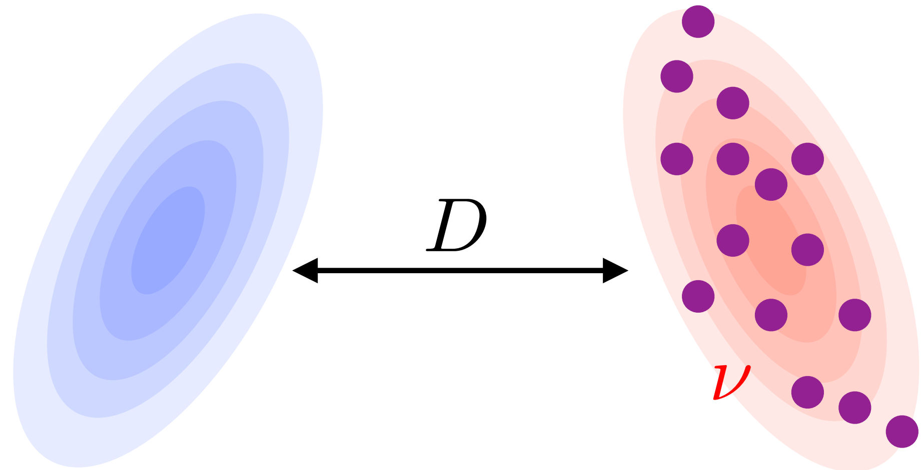
$\Downarrow$

In law

$$\mathbb{P}(X_n \in A) \xrightarrow{n\to+\infty} \mathbb{P}(X \in A)$$

(the X_n can be defined on different spaces)

$$\mu_\theta = f_\theta(x)\mathrm{d}x$$



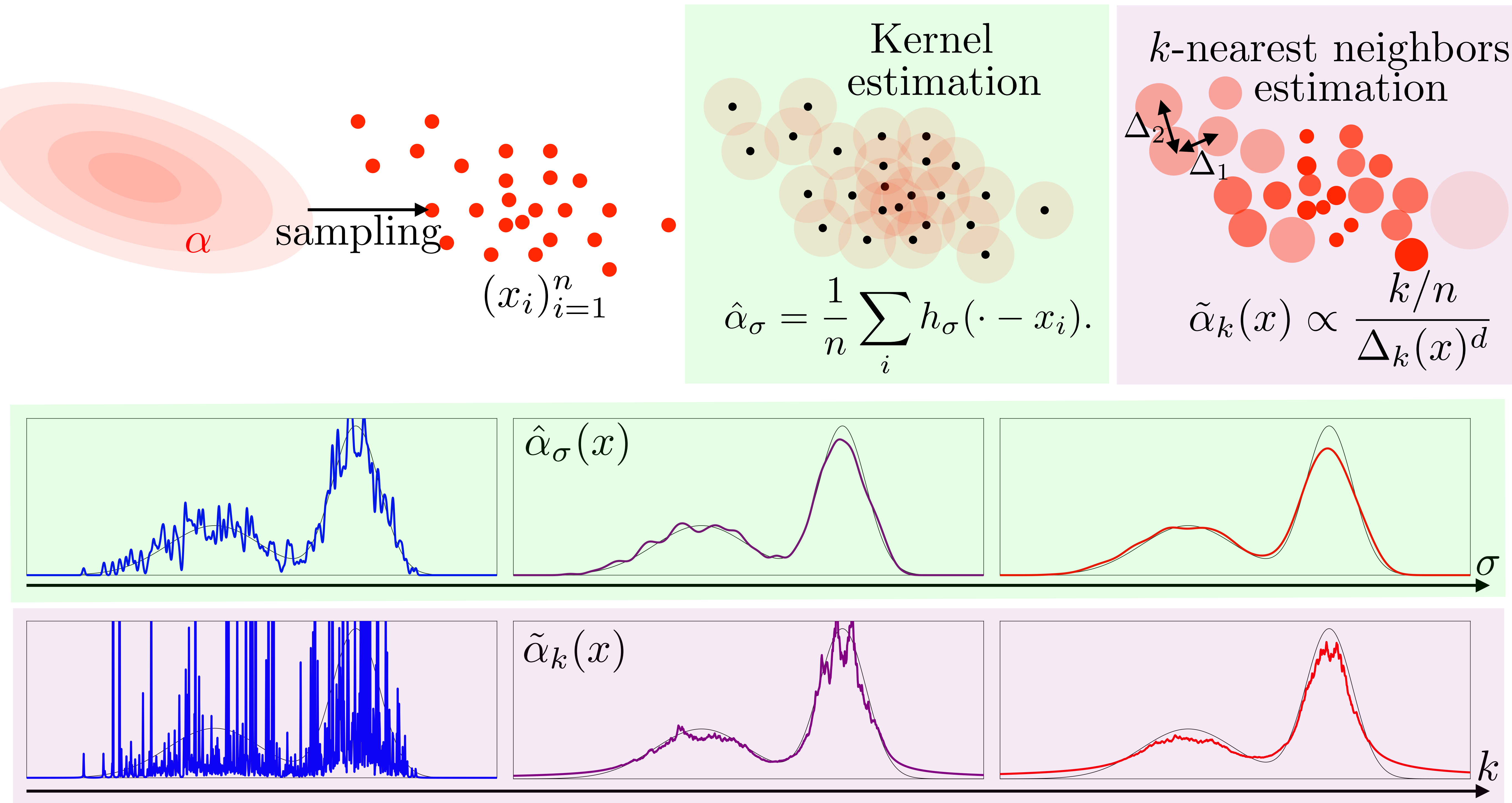
$$\hat{\nu} \stackrel{\text{def.}}{=} \sum_{i=1}^{n} \delta_{x_i}$$

Density fitting: $\quad \min_\theta D(\mu_\theta, \hat{\nu})$

Maximum likelihood estimator

$$\min_\theta -\sum_i \log(f_\theta(x_i)) \xrightarrow{n \to +\infty} \mathrm{KL}(\nu|\mu_\theta) = \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_\theta}\right)\mathrm{d}\nu$$

Minimum Kantorovitch estimator $\quad D = \text{Wasserstein.}$

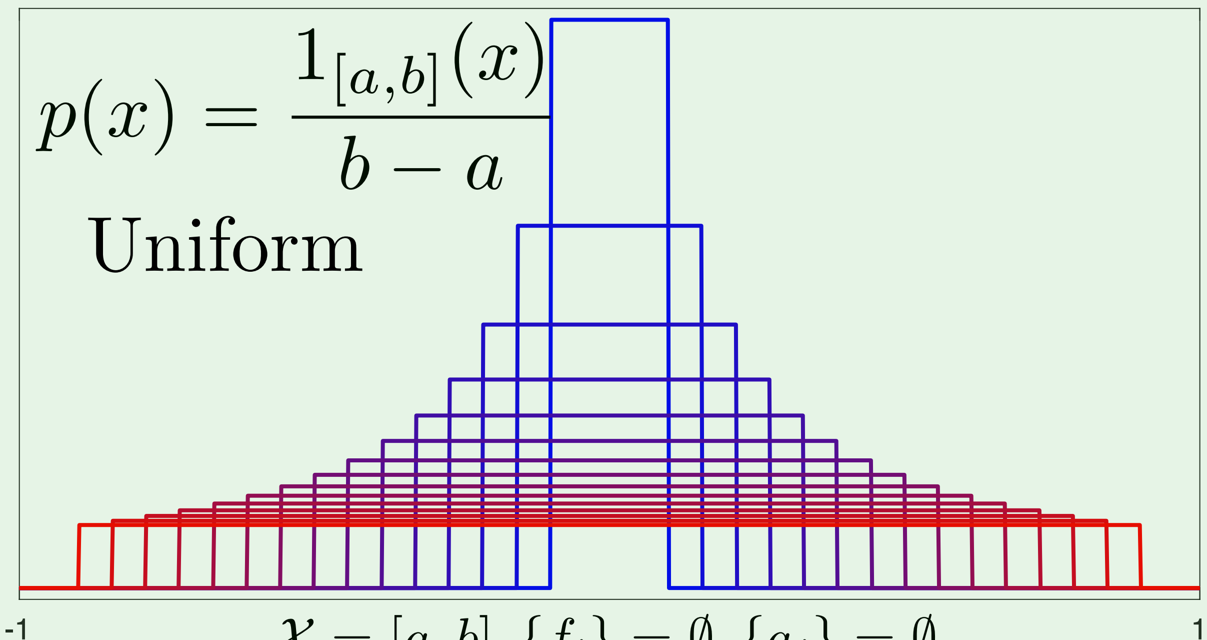


sampling
α
$(x_i)_{i=1}^n$
Kernel estimation
$\hat{\alpha}_\sigma = \frac{1}{n}\sum_i h_\sigma(\cdot - x_i).$
k-nearest neighbors estimation
Δ_2
Δ_1
$\tilde{\alpha}_k(x) \propto \frac{k/n}{\Delta_k(x)^d}$
$\hat{\alpha}_\sigma(x)$
σ
$\tilde{\alpha}_k(x)$
k

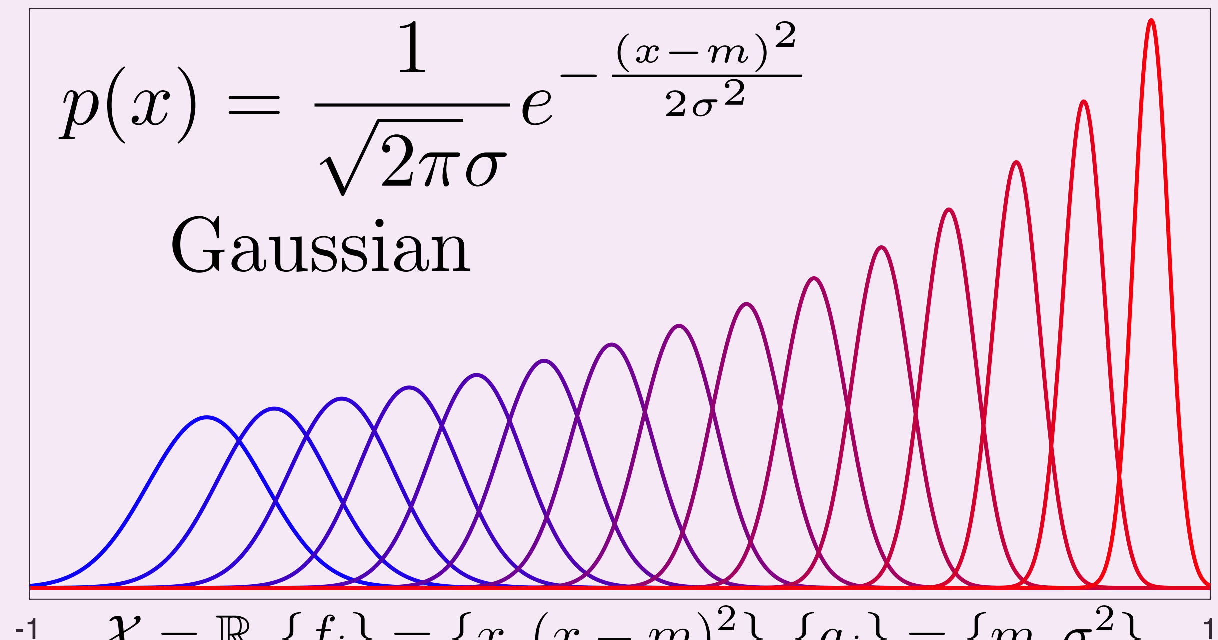
Maximum entropy:

$$\max_p \left\{ -\int_{\mathcal{X}} p(x)\log(p(x))\mathrm{d}x \; ; \; \forall i, \int_{\mathcal{X}} p(x) f_i(x) = a_i \right\}$$

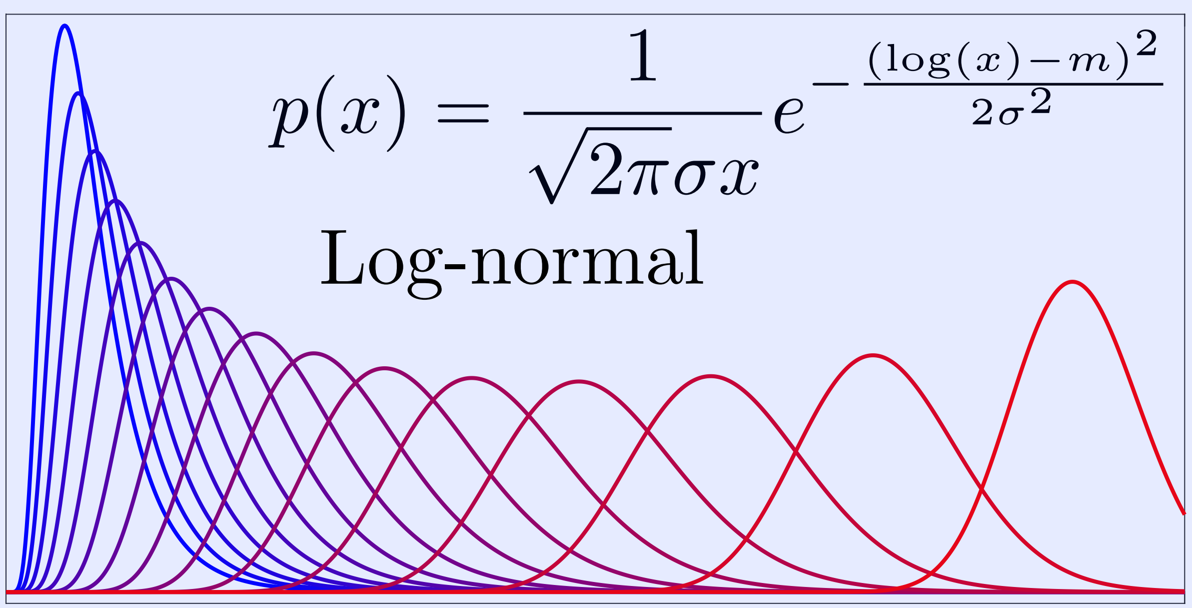
$$\Longrightarrow \quad \exists (\lambda_i)_i, \exists c, \quad p(x) = c\exp\left(\sum_i \lambda_i f_i(x)\right)$$



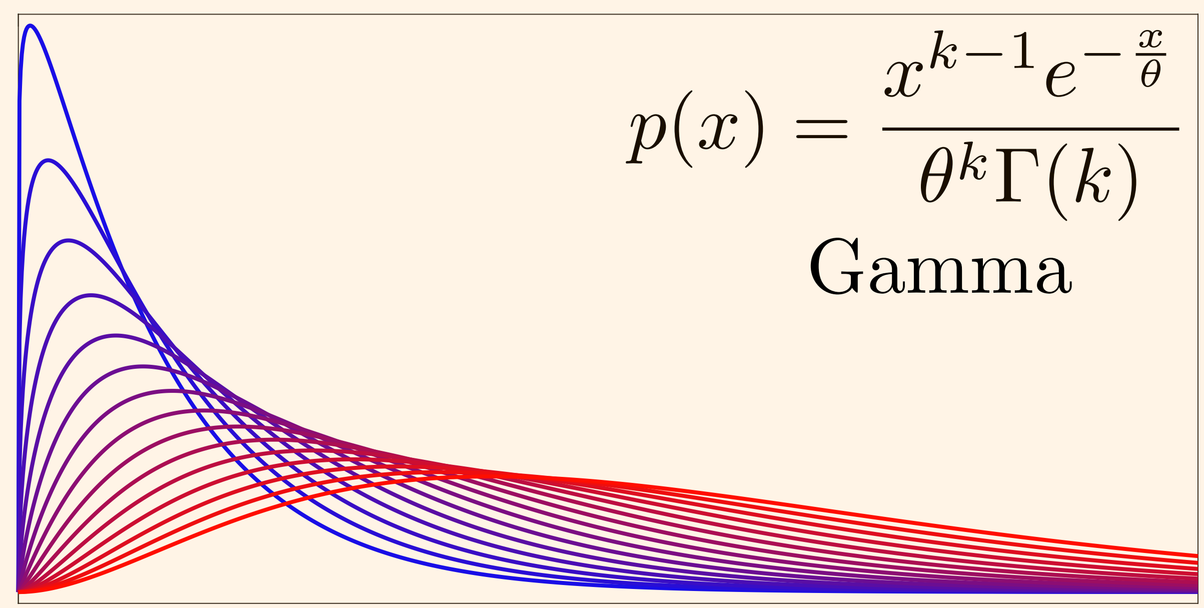
$\mathcal{X} = [a,b], \{f_i\} = \emptyset, \{a_i\} = \emptyset$



$\mathcal{X} = \mathbb{R}, \{f_i\} = \{x, (x-m)^2\}, \{a_i\} = \{m, \sigma^2\}$



$\mathcal{X} = \mathbb{R}^+, \{f_i\} = \{\log(x), (\log(x)-m)^2\}, \{a_i\} = \{m, \sigma^2\}$



$\mathcal{X} = \mathbb{R}^+, \{f_i\} = \{x, \log(x)\}, \{a_i\} = \{k\theta, \psi(k) + \log(\theta)\}$

A frequently asked question by good students is to know if one can replace the convergence in law by the (stronger) convergence in probability. The answer is negative, and in particular the convergence cannot hold almost surely or in L^p. Let us examine why. We proceed by contradiction. Suppose that $Z_n \to Z_\infty$ in probability for a random variable Z_∞ (necessarily of standard Gaussian law). Then on the one hand, by the triangle inequality, for any $\varepsilon > 0$,

$$\mathbb{P}(|Z_{2n} - Z_n| \geq 2\varepsilon) \leq \mathbb{P}(|Z_{2n} - Z_\infty| \geq \varepsilon) + \mathbb{P}(|Z_{2n} - Z_\infty| \geq \varepsilon) \to 0.$$

and therefore $Z_{2n} - Z_n \to 0$ in probability. On the other hand we have

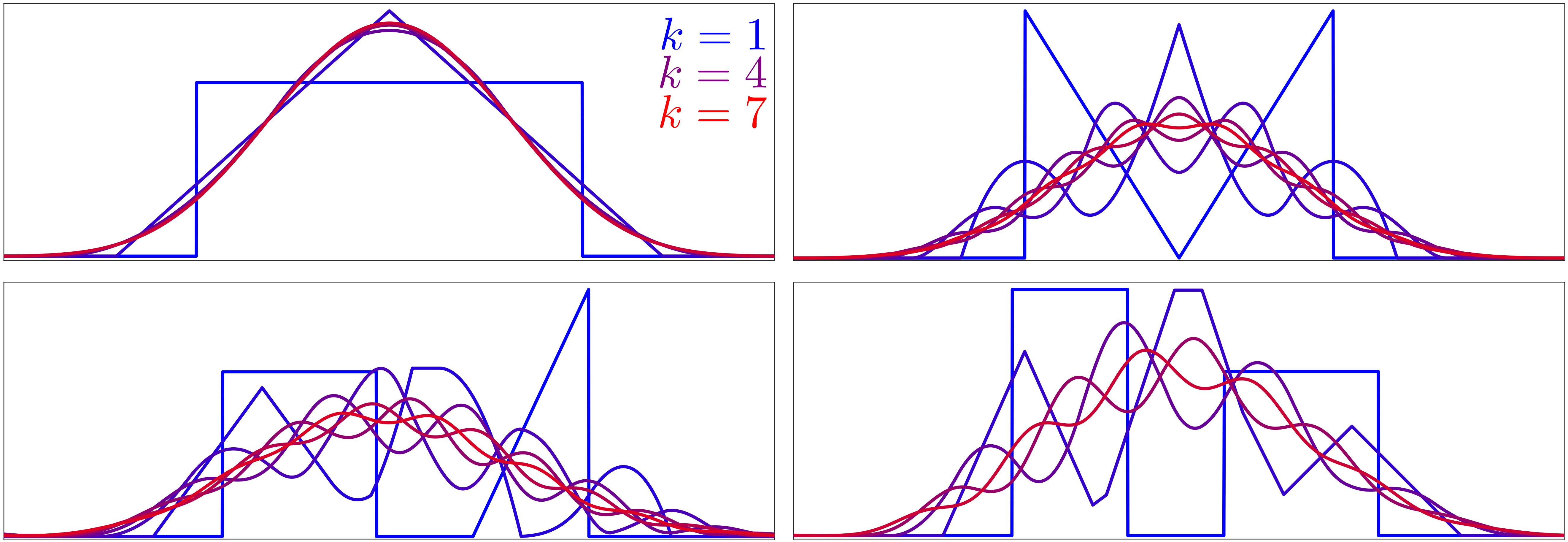
$$\begin{aligned} Z_{2n} - Z_n &= \frac{X_1 + \cdots + X_{2n} - \sqrt{2}(X_1 + \cdots + X_n)}{\sqrt{2n}} \\ &= \frac{1-\sqrt{2}}{\sqrt{2}} Z_n + \frac{X_{n+1} \cdots + X_{2n}}{\sqrt{2n}} \\ &= \frac{1-\sqrt{2}}{\sqrt{2}} Z_n + \frac{1}{\sqrt{2}} Z'_n. \end{aligned}$$

Now Z'_n is an independent copy of Z_n, and therefore the CLT used twice shows that $Z_{2n} - Z_n$ converges in law as $n \to \infty$ to a Gaussian law of mean zero and variance $(1 - \sqrt{2})^2/2 + 1/2 = 2 - \sqrt{2} \neq 0$. Hence the contradiction.

$$\left.\begin{array}{l} Z_n \xrightarrow{\mathbb{P}} Z_\infty \Rightarrow \mathbb{P}(|Z_{2n} - Z_n| \geq 2\varepsilon) \leq \mathbb{P}(|Z_{2n} - Z_\infty| \geq \varepsilon) + \mathbb{P}(|Z_{2n} - Z_\infty| \geq \varepsilon) \to 0. \Rightarrow Z_{2n} - Z_n \xrightarrow{\mathbb{P}} 0 \\ Z_{2n} - Z_n = \frac{X_1 + \cdots + X_{2n} - \sqrt{2}(X_1 + \cdots + X_n)}{\sqrt{2n}} = \frac{1-\sqrt{2}}{\sqrt{2}} Z_n + \frac{1}{\sqrt{2}} Z'_n. \underset{(\text{TCL} \times 2)}{\xrightarrow{\mathbb{P}}} \mathcal{N}(0, 2 - \sqrt{2}) \end{array}\right\} \text{contradiction}$$

Central limit theorem:

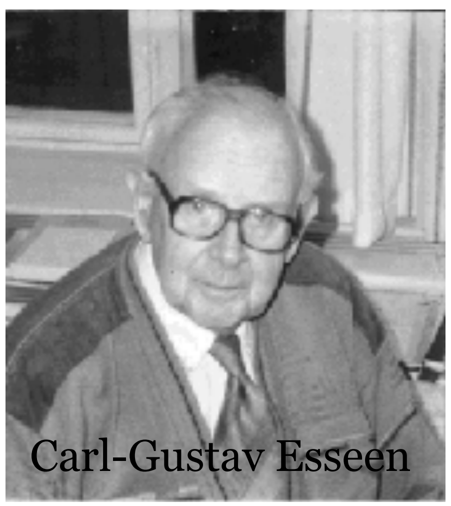
If $\int (1, x, x^2) f(x) \mathrm{d}x = (1, 0, 1)$, then $\underbrace{(f \star \ldots \star f)}_{k \text{ times}} \left(\frac{x}{\sqrt{k}} \right) \xrightarrow[\mathbb{P}]{k \to +\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$



Central limit theorem: If $\mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1$ and $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$

$$Y_n \overset{\text{def.}}{=} \frac{X_1 + \ldots + X_n}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0,1)$$

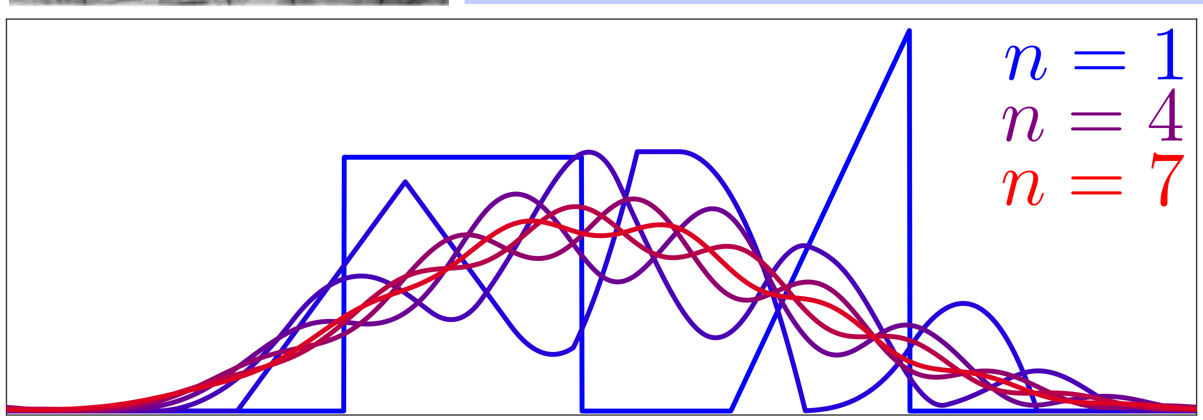
Kolmogorov-Smirnov distance: $d_{KS}(X,Y) \overset{\text{def.}}{=} \max_t |\mathbb{P}(X \leqslant t) - \mathbb{P}(Y \leqslant t)|$

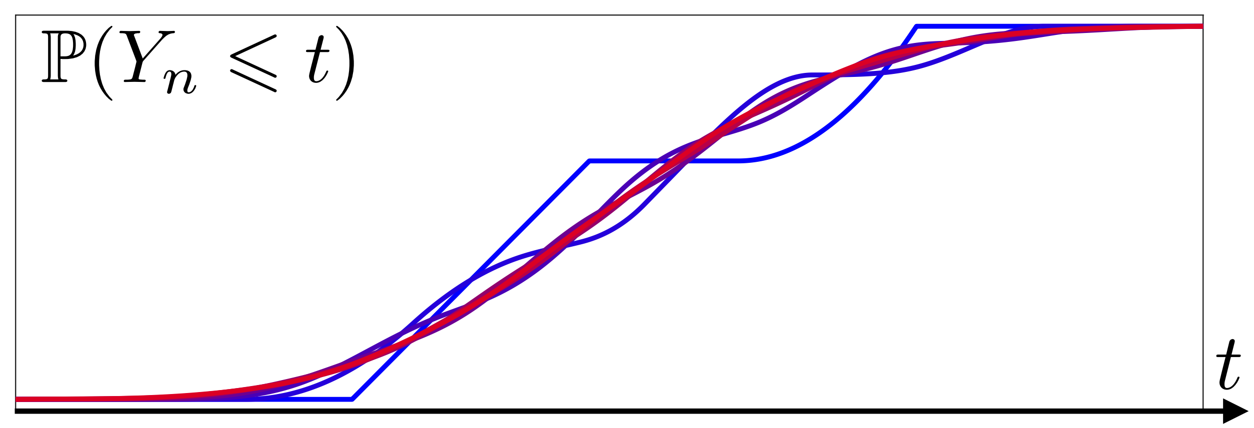


Metrizes convergence in law: $X \overset{\text{law}}{\to} Y \Leftrightarrow d_{\text{KS}}(X,Y) \to 0$

Theorem:
[Berry 1941]
[Esseen, 1942]

$$d_{\text{KS}}(Y_n, \mathcal{N}(0,1)) \leqslant \frac{C\mathbb{E}(|X|^3)}{\sqrt{n}} \qquad C \leqslant 1/2$$



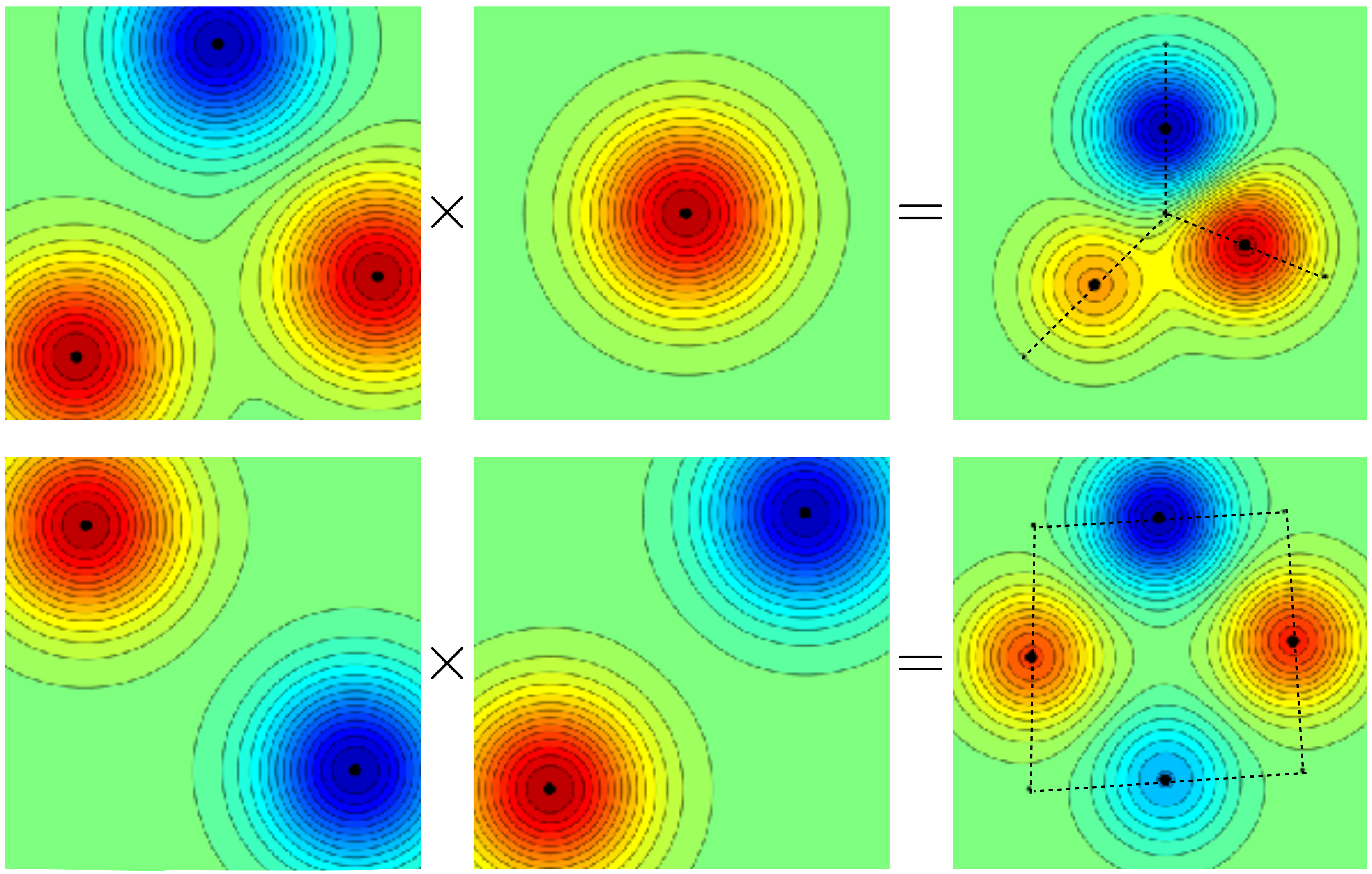


Gaussian bump: $G_{m,s}(x) \overset{\text{def.}}{=} e^{-\frac{\|x-m\|^2}{2s}}$

Prop: $G_{m_0,s_0}(x)G_{m_1,s_1}(x) = \rho G_{m,s}(x)$

$$\frac{1}{s} = \frac{1}{s_0} + \frac{1}{s_1} \qquad \rho \leqslant 1$$

$$m = \frac{s}{s_0}m_0 + \frac{s}{s_1}m_1$$



Hump algebra

Random walk: $x_{k+1} = x_k + \frac{\sigma}{\sqrt{K}} \varepsilon_k$

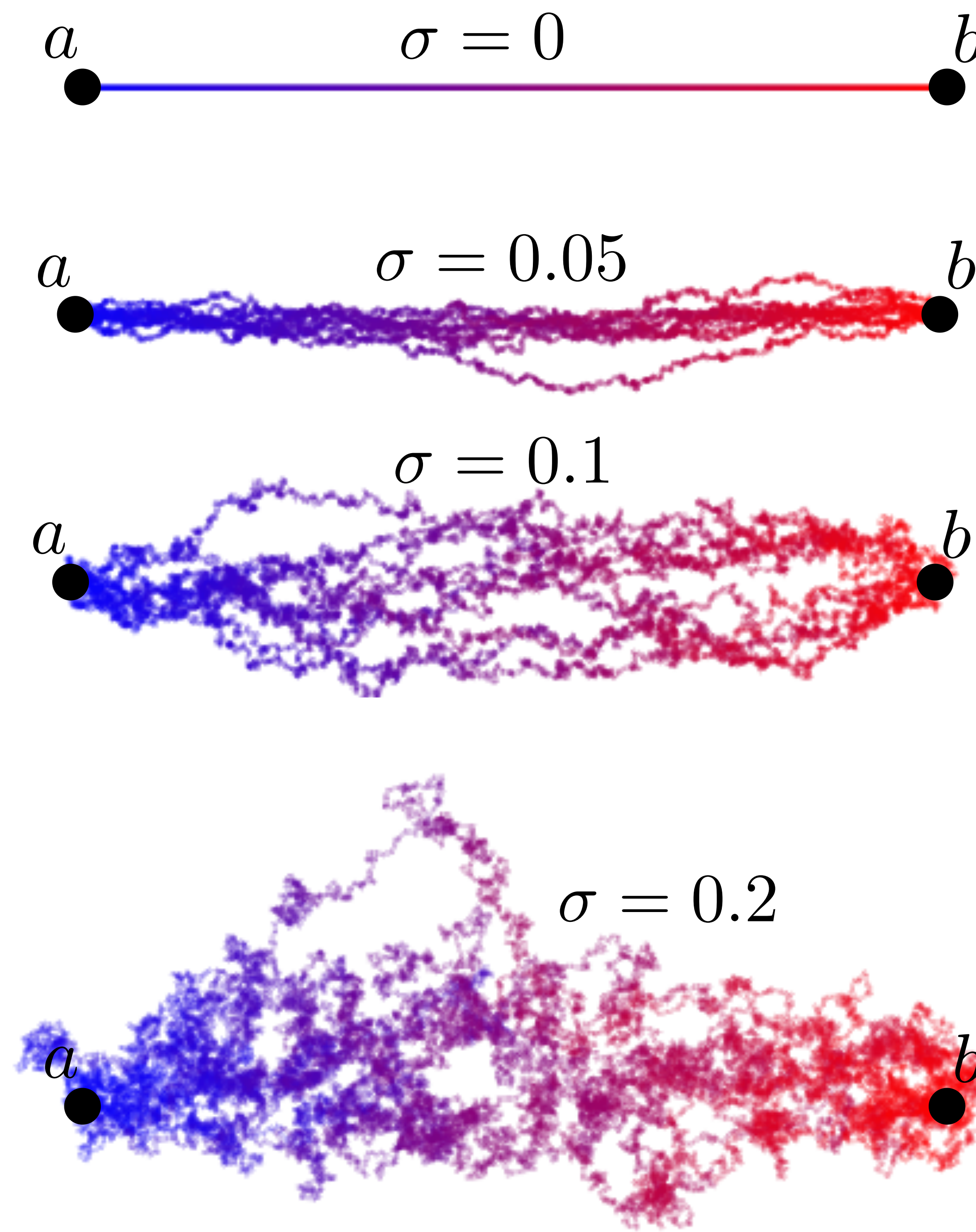
$\varepsilon_k \sim \mathcal{N}(0, \mathrm{Id}_{\mathbb{R}^2})$

Brownian motion / Wiener process:

$$x_k \xrightarrow[k/K \to t]{K \to +\infty} W_t$$

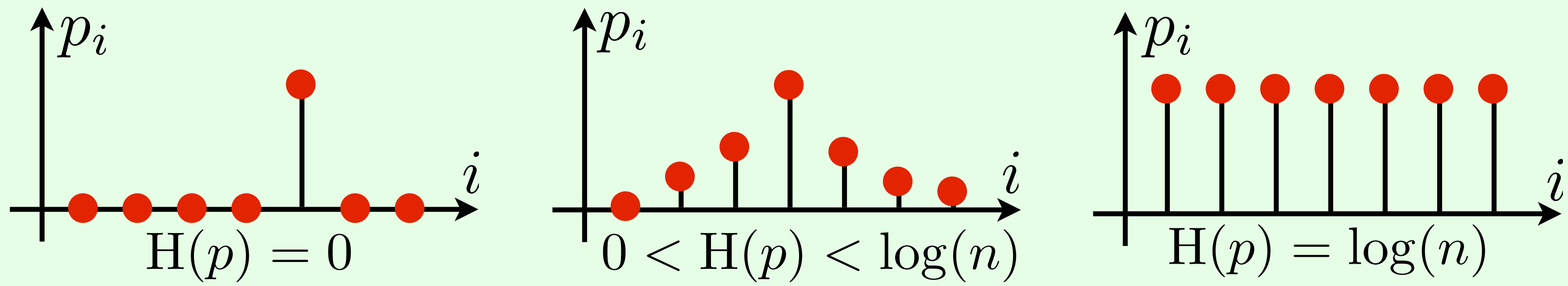
Brownian bridge between $(a, b) \in \mathbb{C}^2$:

$$a + (b - a) \frac{x(t) - x(0)}{x(1) - x(0)}$$



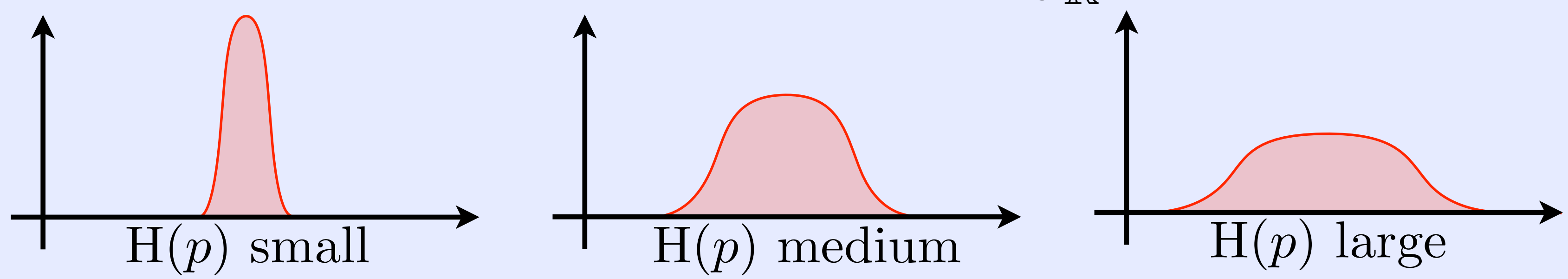
Discrete

$p_i \geqslant 0, \sum_{i=1}^n p_i = 1 \qquad \mathrm{H}(p) \stackrel{\text{def.}}{=} -\sum_i p_i \log(p_i)$



Continuous

$p(x) \geqslant 0, \int_{\mathbb{R}^d} p(x) = 1 \quad \mathrm{H}(p) \stackrel{\text{def.}}{=} -\int_{\mathbb{R}^d} p(x)\log(p(x))\mathrm{d}x$



General

Relative entropy (Kullback-Leibler)

Measures (μ, ν): $\quad \mathrm{KL}(\mu|\nu) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\mu(x)$

$\mathrm{H}(p) = -\mathrm{KL}(p\,\mathrm{d}x|\mathrm{d}x)$

Goal: sample from $\mathbb{P}_X(x) = \frac{1}{Z} f(x)$.

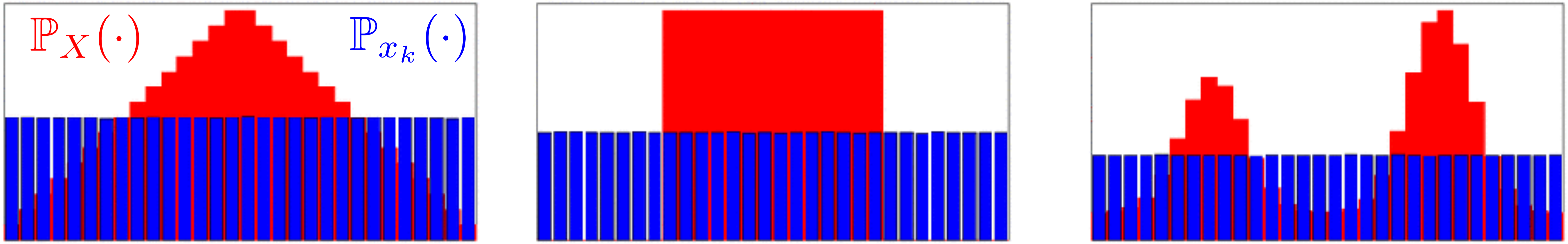
unknown ↗

Needs: transition probability $\mathbb{P}_{Y|X}(y|x)$

$x_0 \leftarrow$ `initialization`

Sample $y_k \sim \mathbb{P}_{Y|X}(\cdot|x_k)$.

$$x_{k+1} \overset{\text{def.}}{=} \begin{cases} x_k & \text{if } \texttt{rand} < \frac{f(y_k)}{f(x_k)} \\ y_k & \text{otherwise.} \end{cases}$$



$\mathbb{P}_{Y|X}(\cdot|x)$ =uniform on neighbors of x.