Batch, Stochastic and Mirror Gradient Descents

Gabriel Peyré

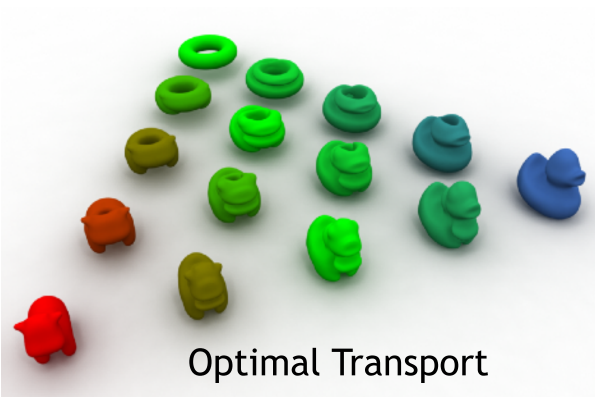
www.numerical-tours.com

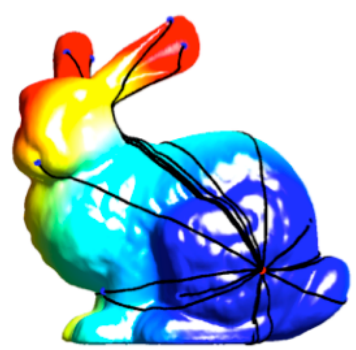
Mathematical Coffees

Huawei-FSMP joint seminars

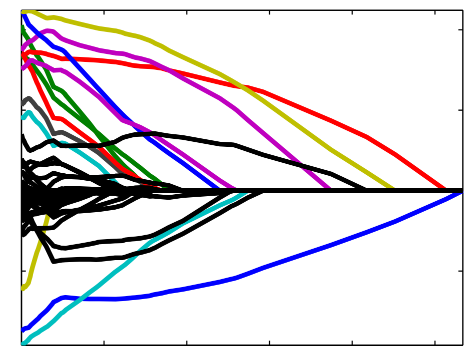
https://mathematical-coffees.github.io

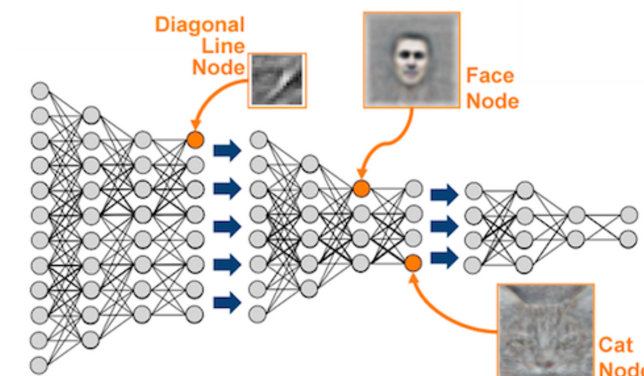
Organized by: Mérouane Debbah & Gabriel Peyré


Optimal Transport

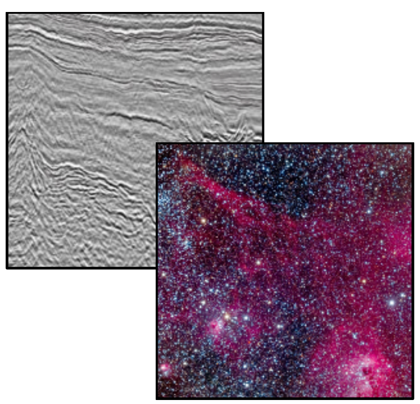

Geodesics

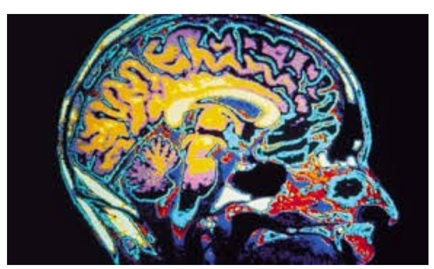

Meshes

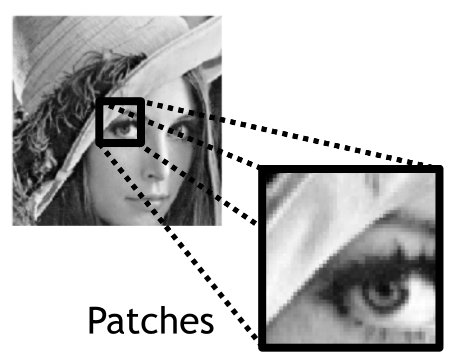

Optimization

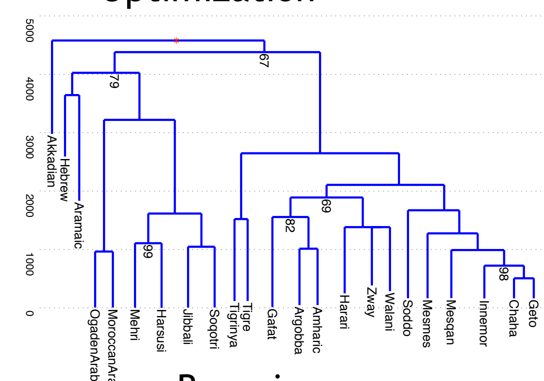


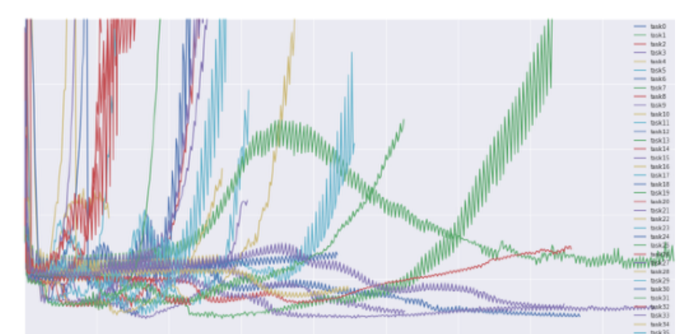
Deep Learning


Sparsity


Neuro-imaging


Patches


Bayesian

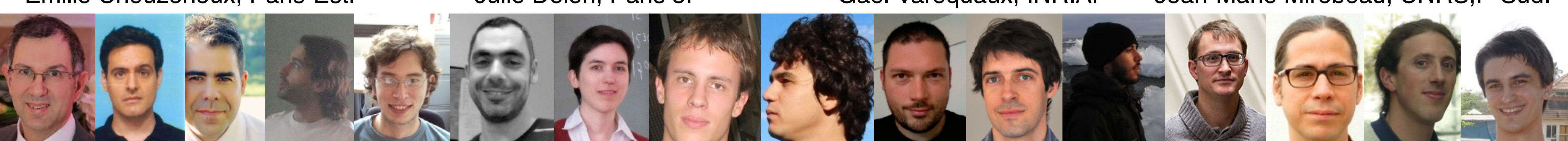

Parallel/Stochastic

Alexandre Allauzen, Paris-Sud.
Pierre Alliez, INRIA.
Guillaume Charpiat, INRIA.
Emilie Chouzenoux, Paris-Est.

Nicolas Courty, IRISA.
Laurent Cohen, CNRS Dauphine.
Marco Cuturi, ENSAE.
Julie Delon, Paris 5.

Fabian Pedregosa, INRIA.
Julien Tierny, CNRS and P6.
Robin Ryder, Paris-Dauphine.
Gael Varoquaux, INRIA.

Jalal Fadili, ENSICaen.
Alexandre Gramfort, INRIA.
Matthieu Kowalski, Supelec.
Jean-Marie Mirebeau, CNRS,P-Sud.



Optimization Everywhere …

Inverse problems: Observations $y = Ax_0 + w$.

Regularized recovery: $\min_x f(x) \overset{\text{def.}}{=} \|y - Ax\|^2 + R(x)$.

Optimization Everywhere ...

Inverse problems: Observations $y = Ax_0 + w$.

Regularized recovery: $\min_x f(x) \overset{\text{def.}}{=} \|y - Ax\|^2 + R(x)$.

Supervised learning: Observations: $(a_i, y_i)_i$, parametric model: $g(x, a)$

Regression:	$y_i \approx g(x, a_i)$	$\ell(y, y') = \lvert y - y' \rvert^2$
Classification:	$y_i \approx \theta(g(x, a_i))$ $\theta(u) = (1 + e^u)^{-1}$	$\ell(y, y') = \log(1 + e^{-yy'})$

Empirical risk minimization: $\min_x f(x) = \frac{1}{n} \sum_i \ell(g(x, a_i), y_i)$

Optimization Everywhere ...

Inverse problems: Observations $y = Ax_0 + w$.

Regularized recovery: $\min_x f(x) \overset{\text{def.}}{=} \|y - Ax\|^2 + R(x)$.

Supervised learning: Observations: $(a_i, y_i)_i$, parametric model: $g(x, a)$

Regression:	$y_i \approx g(x, a_i)$	$\ell(y, y') = \|y - y'\|^2$
Classification:	$y_i \approx \theta(g(x, a_i))$ $\theta(u) = (1 + e^u)^{-1}$	$\ell(y, y') = \log(1 + e^{-yy'})$

Empirical risk minimization: $\min_x f(x) = \frac{1}{n} \sum_i \ell(g(x, a_i), y_i)$

$$\min_x f(x)$$

finite sum / empirical		integral / expectation
$f(x) \overset{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n f_i(x)$	← sampling $n \to +\infty$ →	$f(x) \overset{\text{def.}}{=} \mathbb{E}_{\mathbf{z}}(f(x, \mathbf{z}))$

Batch Gradient Descent

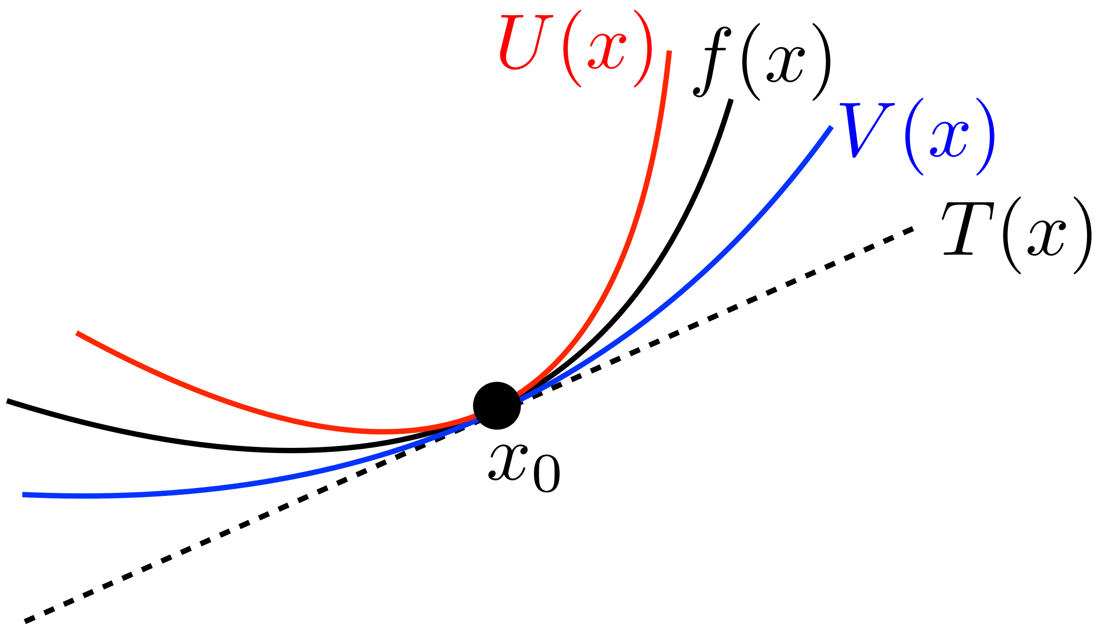
$$x_{k+1} = x_k - \tau_k \nabla f(x_k)$$

Hypotheses: $\mu \mathrm{Id}_n \preceq \partial^2 f(x) \preceq L\, \mathrm{Id}_n$

strong convexity smoothness

Conditionning:

$$\varepsilon \overset{\text{def.}}{=} \frac{L}{\mu} \leqslant 1$$



$$T(x) \overset{\text{def.}}{=} f(x_0) + \langle \nabla f(x_0),\, x - x_0 \rangle$$

$$U(x) \overset{\text{def.}}{=} T(x) + \frac{L}{2}\|x - x_0\|^2$$

$$V(x) \overset{\text{def.}}{=} T(x) + \frac{\mu}{2}\|x - x_0\|^2$$

$$\Rightarrow \|x - x^\star\|^2 \leqslant \frac{f(x_0) - f(x^\star)}{\mu/2}$$

Batch Gradient Descent

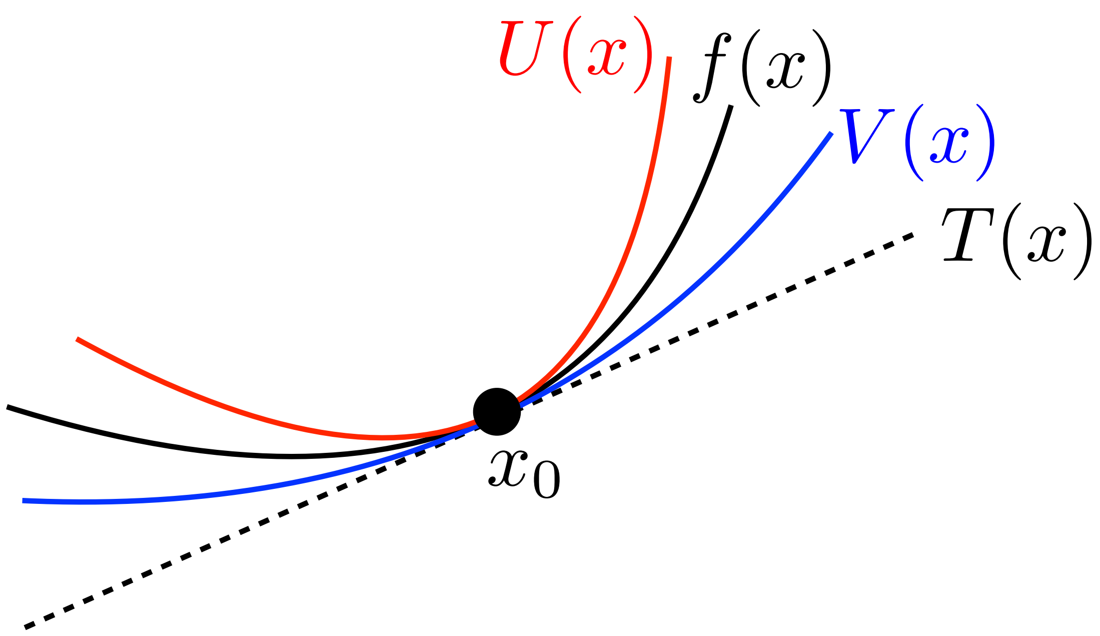
$$x_{k+1} = x_k - \tau_k \nabla f(x_k)$$

Hypotheses: $\mu \mathrm{Id}_n \preceq \partial^2 f(x) \preceq L\, \mathrm{Id}_n$

strong convexity — smoothness

Conditionning:

$$\varepsilon \overset{\text{def.}}{=} \frac{L}{\mu} \leqslant 1$$



$$T(x) \overset{\text{def.}}{=} f(x_0) + \langle \nabla f(x_0),\, x - x_0 \rangle$$

$$U(x) \overset{\text{def.}}{=} T(x) + \frac{L}{2}\|x - x_0\|^2$$

$$V(x) \overset{\text{def.}}{=} T(x) + \frac{\mu}{2}\|x - x_0\|^2$$

$$\Rightarrow \|x - x^\star\|^2 \leqslant \frac{f(x_0) - f(x^\star)}{\mu/2}$$

Theorem:

If $L < +\infty$, $0 < \tau < \frac{2}{L}$ $\qquad f(x_k) - f(x^\star) \leqslant \frac{C}{\ell + 1}$

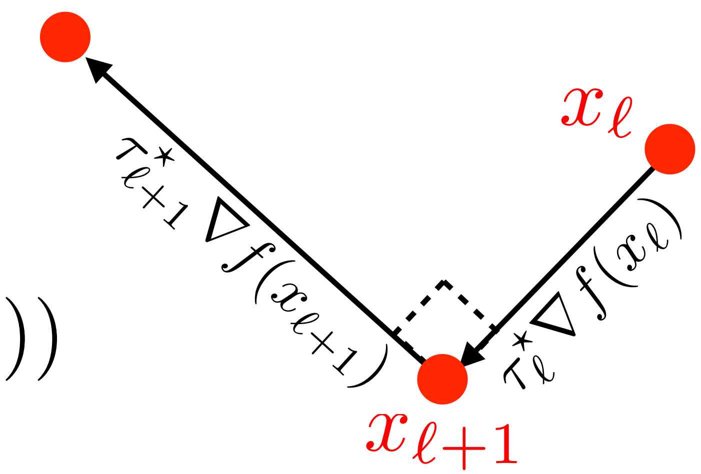
If $\mu > 0$, $L < +\infty$, $0 < \tau < \frac{2}{L}$ $\qquad \|x_k - x^\star\| \leqslant \rho^\ell \|x_0 - x^\star\|$

$$\rho = (1 + \varepsilon)^{-\frac{1}{2}} < 1$$

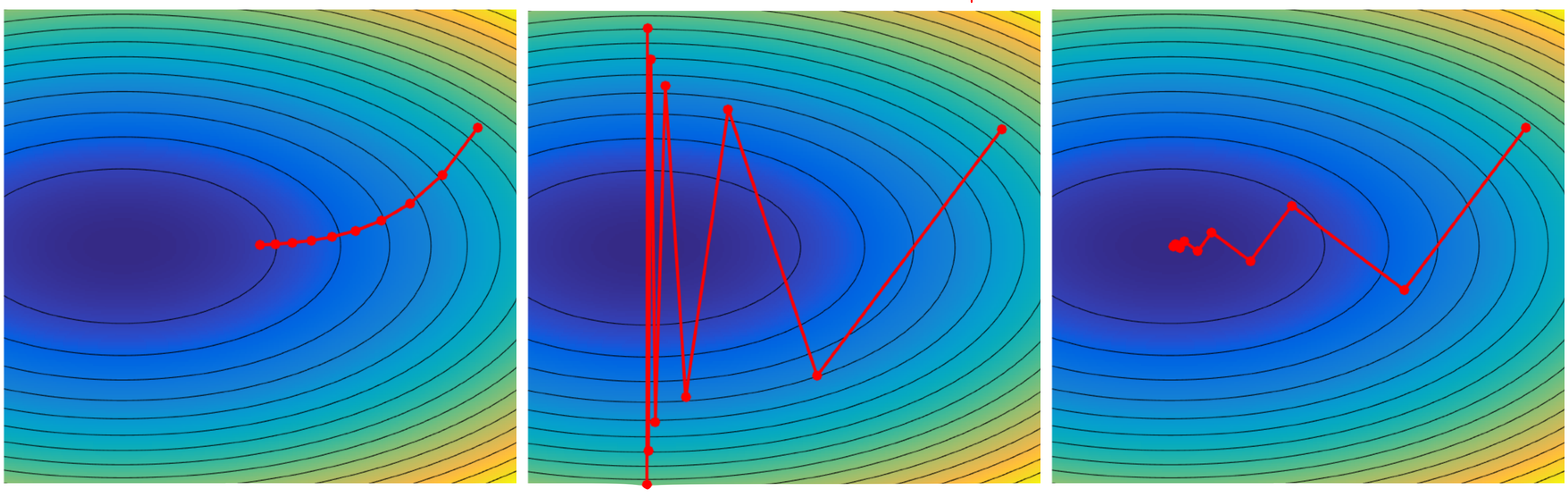
Step size matters ...

$$x_{\ell+1} = x_\ell - \tau_\ell \nabla f(x_\ell)$$

$$\tau_\ell^\star = \underset{\tau}{\operatorname{argmin}}\ f(x_\ell - \tau \nabla f(x_\ell))$$



$$\nabla f(x_\ell) \perp \nabla f(x_{\ell+1})$$



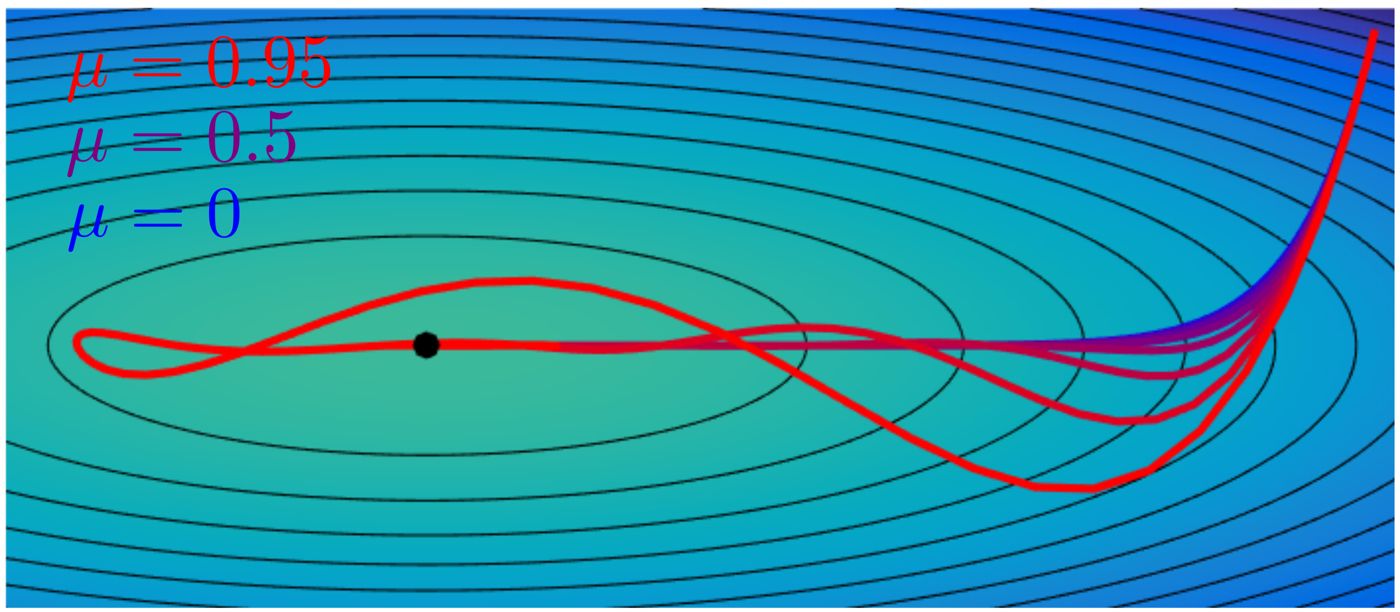
Small τ_ℓ | Large τ_ℓ | Optimal $\tau_\ell = \tau_\ell^\star$

Acceleration

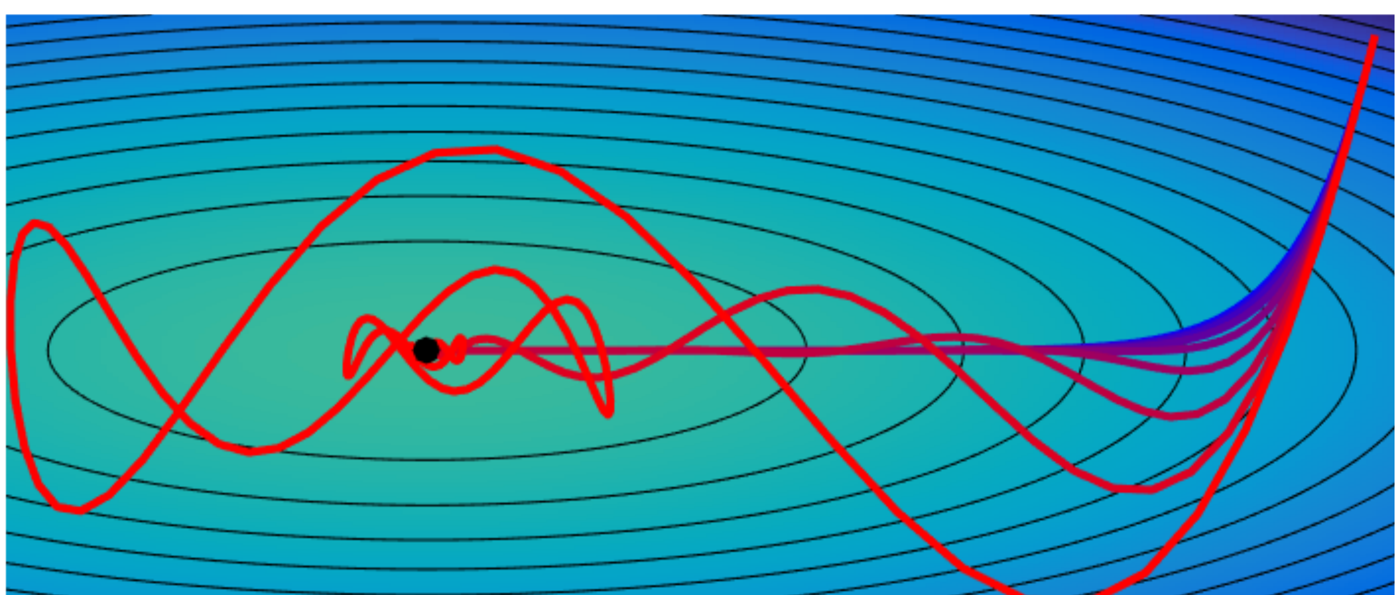
Momentum
"heavy ball"

$$x_{k+1} = x_k + p_k$$
$$p_{k+1} = \mu_k p_k - \tau \begin{cases} \nabla f(x_k) & \text{Polyak} \\ \nabla f(x_k + \mu_k p_k) & \text{Nesterov} \end{cases}$$

Yurii Nesterov



Boris Polyak



Theorem: [Nesterov]

For $\mu_k = \frac{k}{k+3}$, then

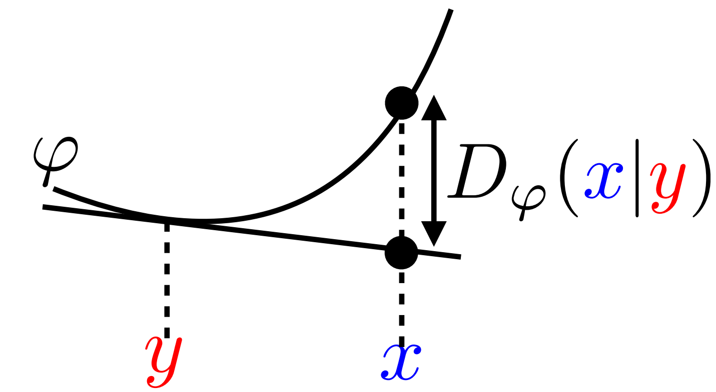
$$f(x_k) - f(x^\star) = O(1/k^2)$$

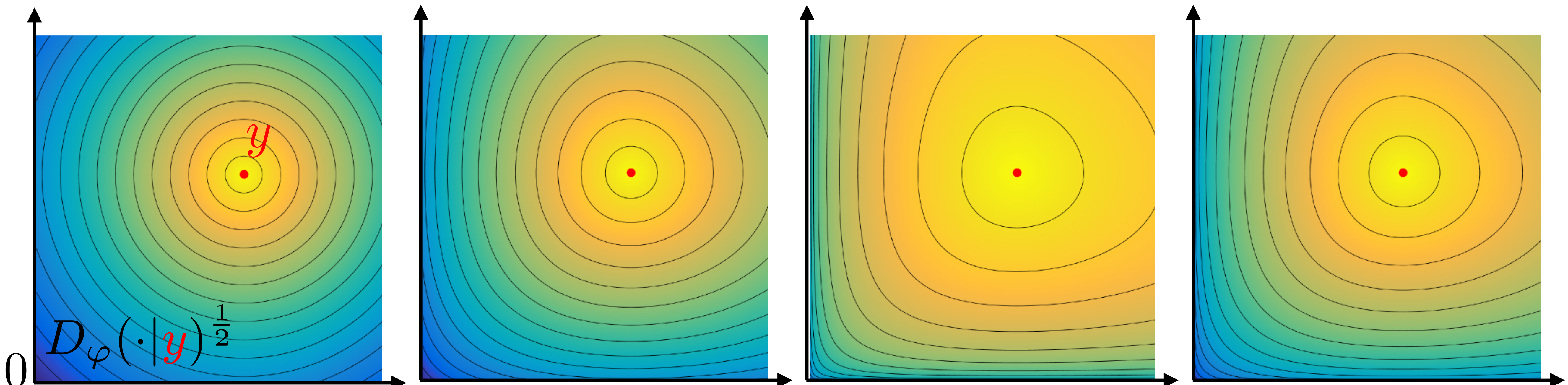
$\rightarrow$ "optimal" for first order schemes.

Generalization: Bregman Divergence

Bregman divergence:

$$D_\varphi(x|y) \stackrel{\text{def.}}{=} \varphi(x) - \varphi(y) - \langle x - y, \nabla\varphi(y)\rangle$$



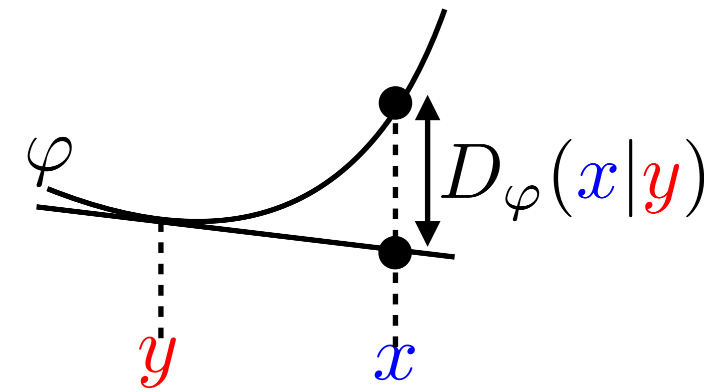


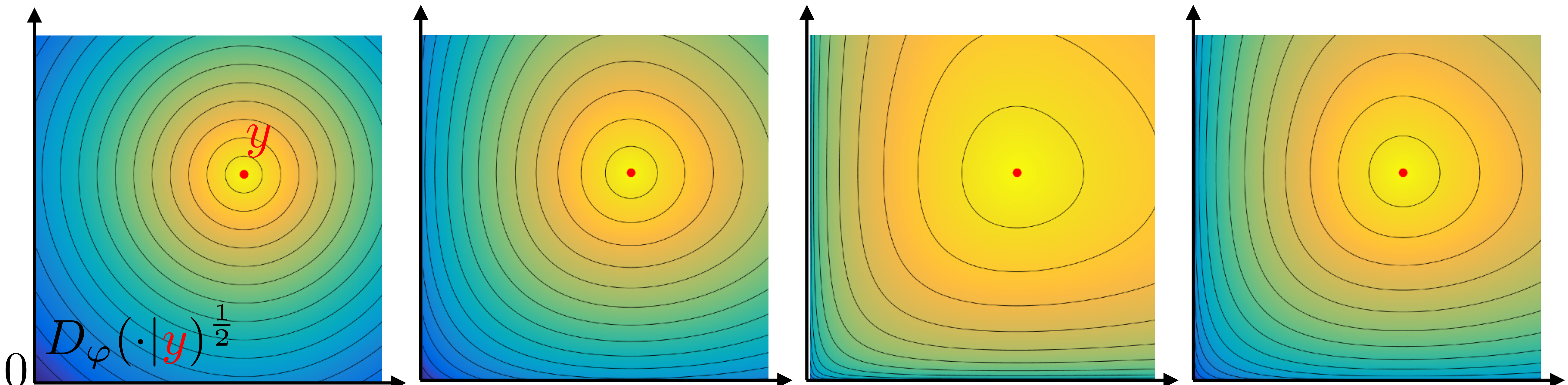
$\varphi(x) = \|x\|^2$ | $\varphi(x) = \sum_i x_i \log(x_i)$ | $\varphi(x) = \sum_i -\log(x_i)$ | $\varphi(x) = \sum_i -\sqrt{x_i}$

Generalization: Bregman Divergence

Bregman divergence:

$$D_\varphi(x|y) \stackrel{\text{def.}}{=} \varphi(x) - \varphi(y) - \langle x - y,\, \nabla\varphi(y)\rangle$$





$\varphi(x) = \|x\|^2$ $\quad$ $\varphi(x) = \sum_i x_i \log(x_i)$ $\quad$ $\varphi(x) = \sum_i -\log(x_i)$ $\quad$ $\varphi(x) = \sum_i -\sqrt{x_i}$

Locally Euclidean:

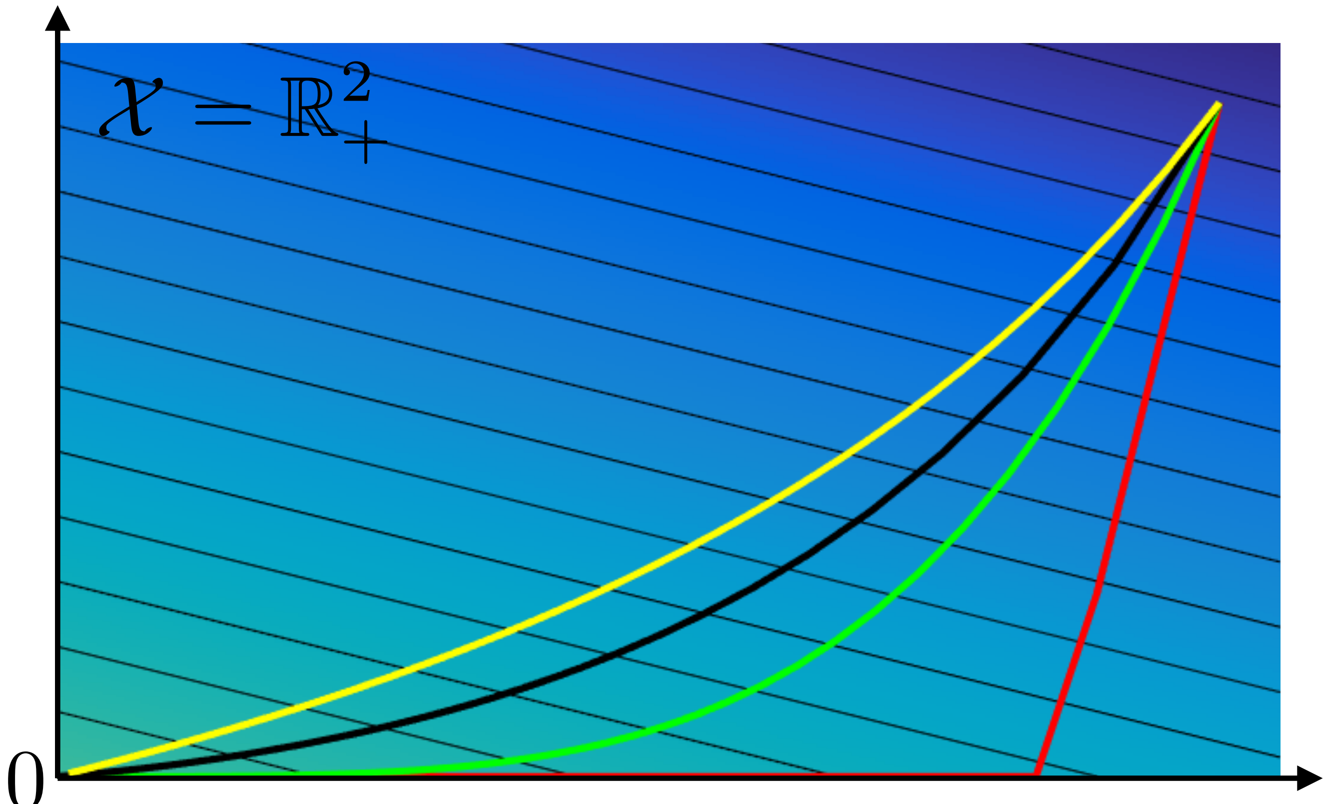
$$D_\varphi(x+\eta|x+\varepsilon) = \frac{1}{2}\langle \partial^2\varphi(x)(\varepsilon-\eta),\, \varepsilon-\eta\rangle + o(\|\varepsilon-\eta\|^2)$$

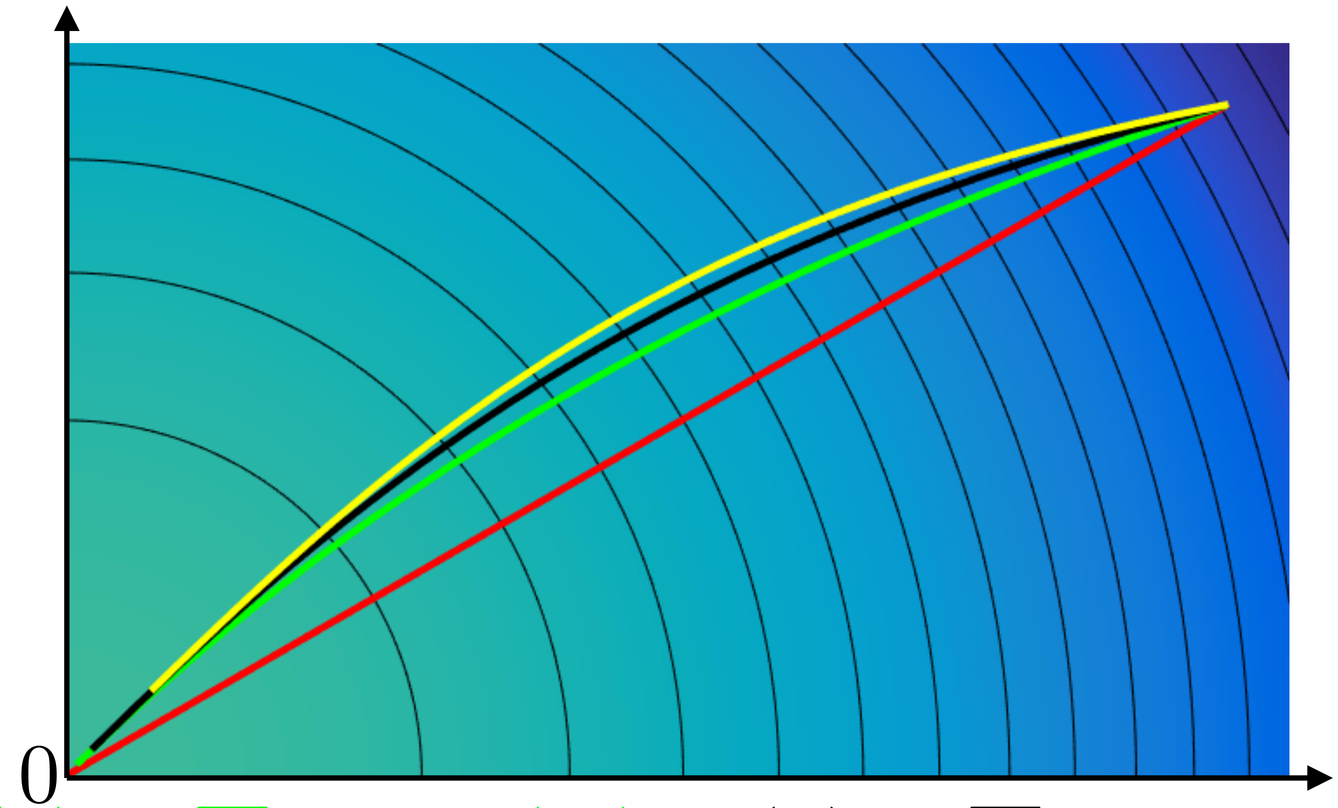
"Rule of thumb:" any reasonnable Euclidean algorithm generalizes to Bregman divergences.

Example: Mirror Descent

Bregman divergence: $D_\varphi(x|y) \stackrel{\text{def.}}{=} \varphi(x) - \varphi(y) - \langle x - y,\, \nabla\varphi(y)\rangle$

Mirror descent:
$$x_{k+1} = \underset{x \in \mathcal{X}}{\text{argmin}}\; D_\varphi(x|x_k) + \tau\langle \nabla f(x_k),\, x\rangle$$
$$= (\nabla\varphi)^{-1}\left(\nabla\varphi(x_k) - \tau\nabla f(x_k)\right)$$

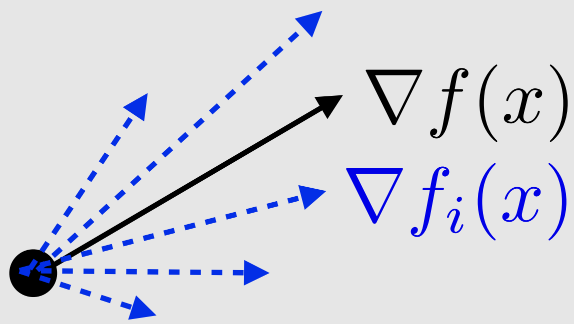




$\varphi(x) = \|x\|^2$ $\varphi(x) = \sum_i -\log(x_i)$ $\varphi(x) = \sum_i x_i \log(x_i)$ $\varphi(x) = \sum_i -\sqrt{x_i}$

Stochastic Gradient Descent

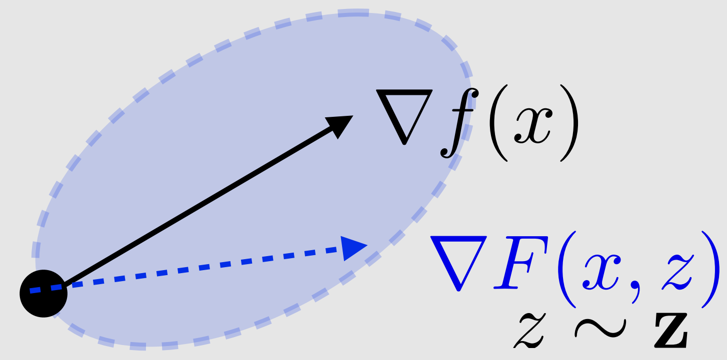
$$f(x) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x)$$
$$\nabla f(x) = \frac{1}{n}\sum_i \nabla f_i(x)$$



Draw $i \in \{1, \ldots, n\}$ uniformly.

$$x_{k+1} = x_k - \tau_k \nabla f_i(x_k)$$

$$f(x) \overset{\text{def.}}{=} \mathbb{E}_{\mathbf{z}}(f(x, \mathbf{z}))$$
$$\nabla f(x) \overset{\text{def.}}{=} \mathbb{E}_{\mathbf{z}}(\nabla F(x, \mathbf{z}))$$



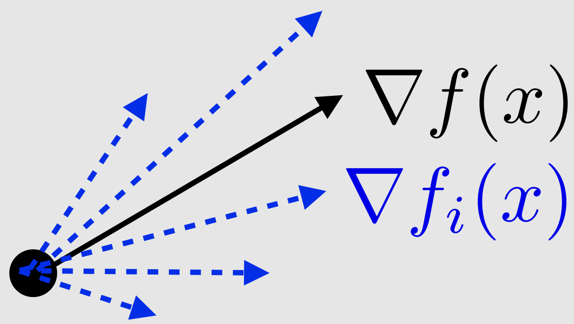
Draw $z \sim \mathbf{z}$

$$x_{k+1} = x_k - \tau_k \nabla F(x, z)$$

Stochastic Gradient Descent

$$f(x) \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

$$\nabla f(x) = \frac{1}{n}\sum_i \nabla f_i(x)$$

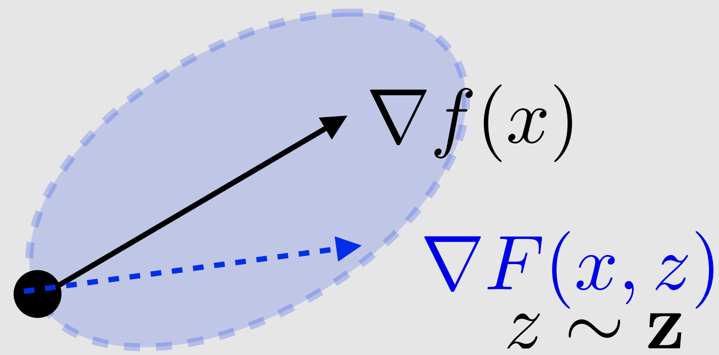


Draw $i \in \{1,\ldots,n\}$ uniformly.

$$x_{k+1} = x_k - \tau_k \nabla f_i(x_k)$$

$$f(x) \stackrel{\text{def.}}{=} \mathbb{E}_{\mathbf{z}}(f(x,\mathbf{z}))$$

$$\nabla f(x) \stackrel{\text{def.}}{=} \mathbb{E}_{\mathbf{z}}(\nabla F(x,\mathbf{z}))$$



Draw $z \sim \mathbf{z}$

$$x_{k+1} = x_k - \tau_k \nabla F(x,z)$$

Theorem: If $\mu > 0$ and $\|\nabla f_i(x)\| \leqslant C$, then for $\tau_k = \frac{1}{\mu(k+1)}$,

$$\mathbb{E}(\|x_k - x^\star\|^2) \leqslant \frac{R}{k+1} \quad \text{where} \quad R \stackrel{\text{def.}}{=} \max(\|x_0 - x^\star\|^2, C^2/\mu^2)$$

$\tau_k \to 0$ to cancel gradient noise.

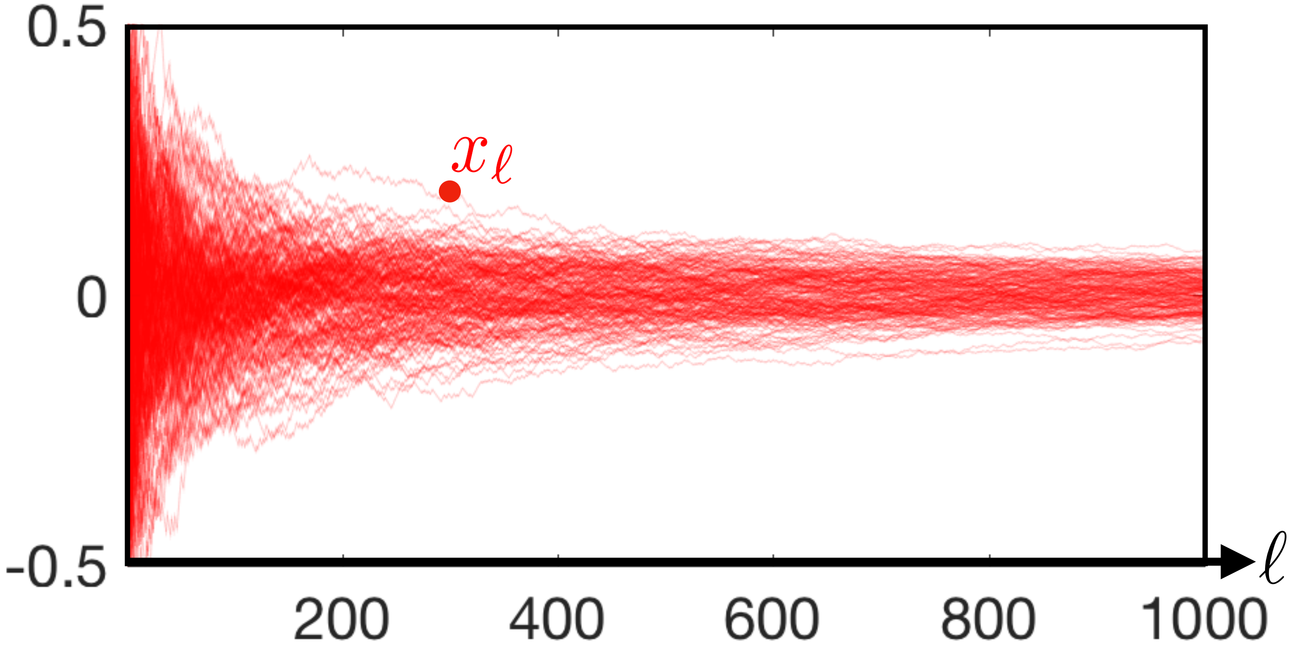
No benefit from strong convexity.

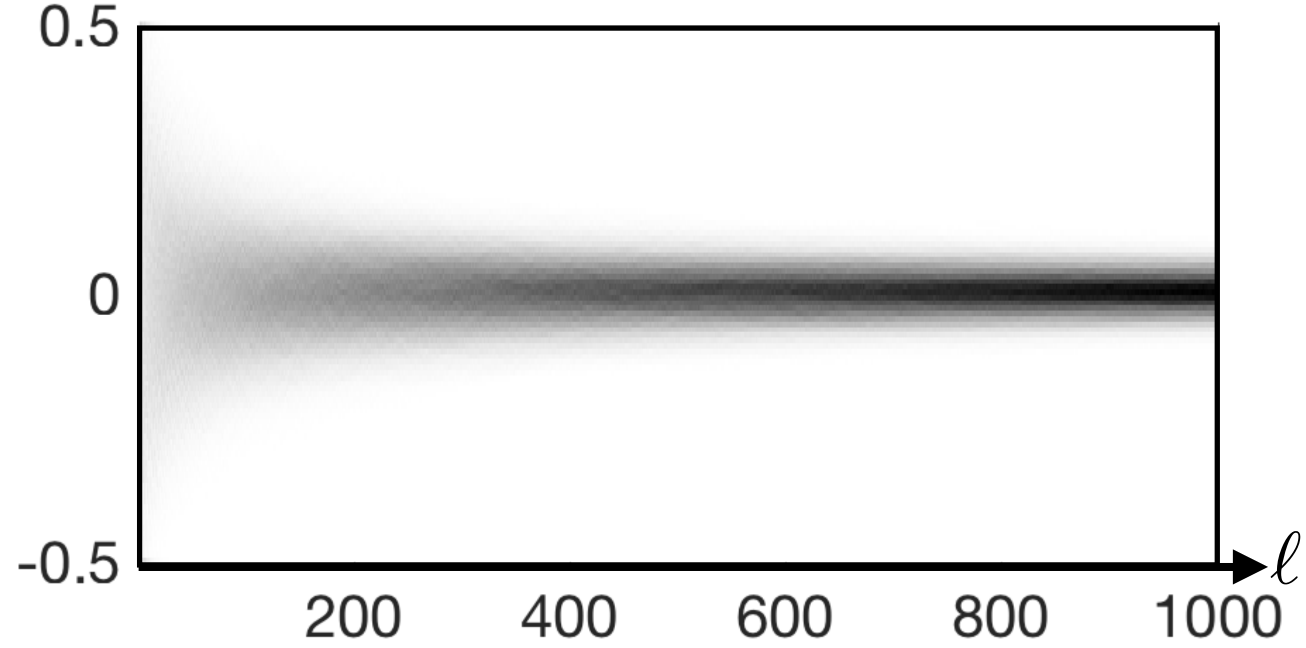
$\longrightarrow$ Only useful when n is *very* large.

Simple Example

$$\min_{x \in \mathbb{R}} \underbrace{(x+1)^2}_{= f_1(x)} + \underbrace{(x-1)^2}_{= f_2(x)}$$

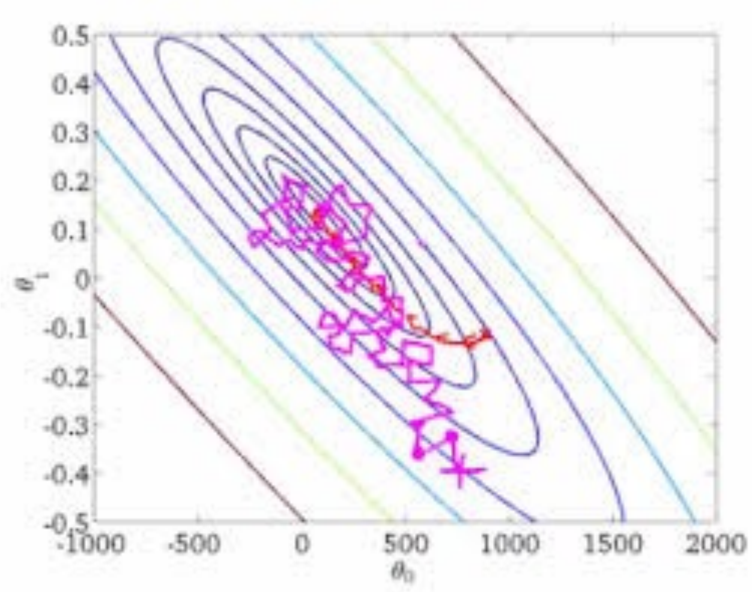
$$x_{\ell+1} \overset{\text{def.}}{=} \begin{cases} x_\ell - \frac{1}{\ell} \nabla f_1(x_\ell) \text{ with proba } \frac{1}{2} \\ x_\ell - \frac{1}{\ell} \nabla f_2(x_\ell) \text{ with proba } \frac{1}{2} \end{cases}$$

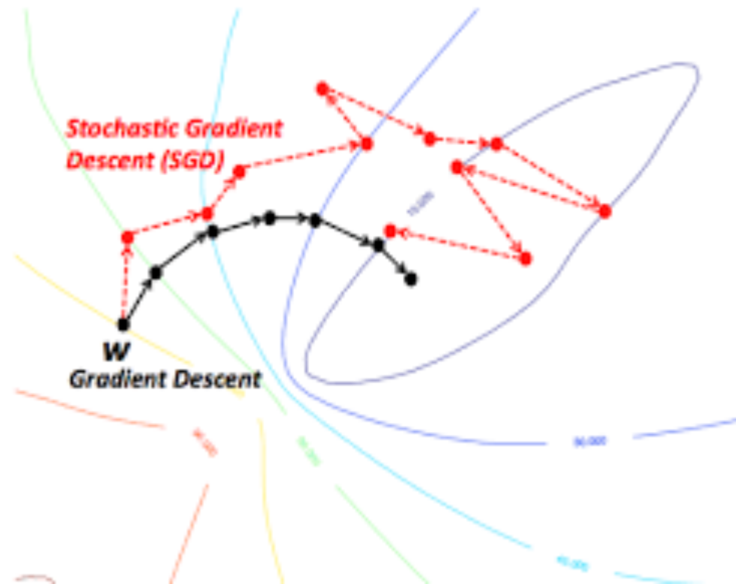


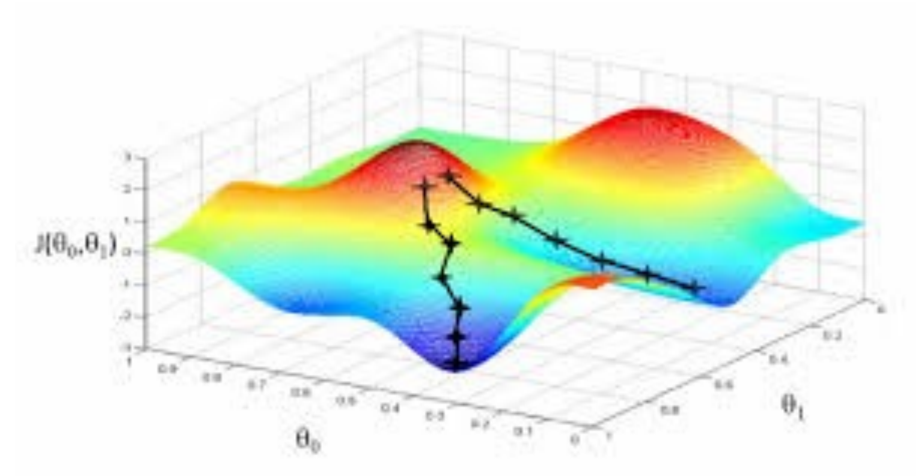


What's Next

Emilie Chouzenoux: stochastic optimization.







Fabian Pedregosa: parallel and distributed optimization.