Around distributional robustness and generalization

in stochastic optimization & machine learning

Franck IUTZELER

Université de Toulouse, Institut de Mathématiques de Toulouse

PGMO Days 2026 · EDF Lab Paris-Saclay · November 18th, 2026

Université de Toulouse

Context: decision under uncertainty

Stochastic programming

decision θ: power production plan, order quantity
uncertainty ξ: demand, renewables, prices
cost $\ell(\theta, \xi)$: operating cost

Machine learning

model θ: parameters of a model, neural network
data point $\xi = (x, y)$: an image x and its label y
loss $\ell(\theta, \xi)$: prediction error

- In both cases, the goal is to minimize the **expected cost**, or true/population risk,

$$\min_{\theta} R(\theta) := \mathbb{E}_{\xi \sim P}\big[\ell(\theta, \xi)\big],$$

  but the probability distribution P is only known through **samples** $\xi_1, \dots, \xi_n$

- Standard proxy: minimizing the **average cost on the data samples**
  sample average approximation (SAA) in stochastic programming
  empirical risk minimization (ERM) in machine learning

$$\min_{\theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, \xi_i)$$

Motivation

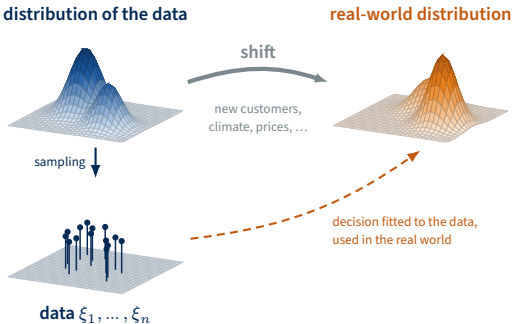
The **traditional approach** of ***"optimizing the average cost"***
is an obstacle to understanding the **robustness** of decisions and models learned from data

Motivation

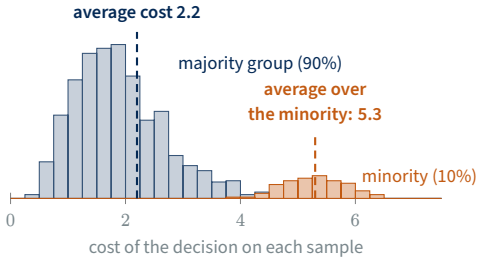
The **traditional approach** of ***"optimizing the average cost"***
is an obstacle to understanding the **robustness** of decisions and models learned from data

I ***The average cost on the data*** is ***not representative*** of real-world performance

Overconfident decisions



Hidden biases



costs of one decision over a dataset with two groups

Overconfidence example: the optimizer's curse

- $\xi_1, \ldots, \xi_n \overset{\text{iid}}{\sim} P$, SAA/ERM solution $\hat{\theta}_n \in \arg\min_\theta \widehat{R}_n(\theta)$

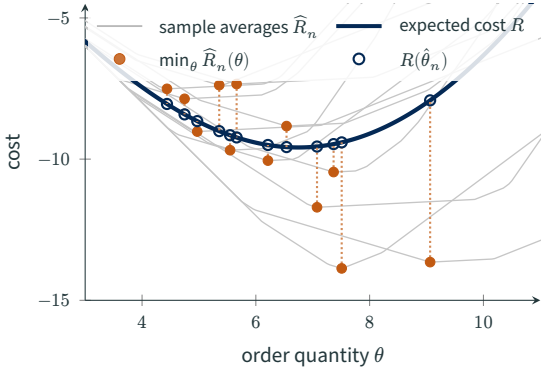
- Since $\mathbb{E}[\widehat{R}_n(\theta)] = R(\theta)$ for every fixed θ,

$$\mathbb{E}\Big[\min_\theta \widehat{R}_n(\theta)\Big] \;\le\; \min_\theta R(\theta) \;\le\; \mathbb{E}[R(\hat{\theta}_n)]$$

- **In the example**, on average over 200 000 datasets, $-10.4 < -9.6 < -8.9$

> The cost of $\hat{\theta}_n$ on the data is an **optimistic estimate** of its cost on new data

Newsvendor $\ell(\theta, \xi) = 5\theta - 7\min(\theta, \xi)$ with Gamma-distributed demand and $n = 10$



12 datasets of $n = 10$ demands

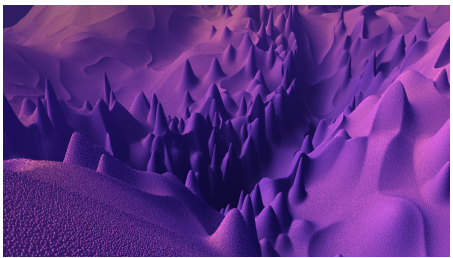
[Mak, Morton & Wood'99] [Smith & Winkler'06] [Shapiro, Dentcheva & Ruszczyński'21]

Motivation

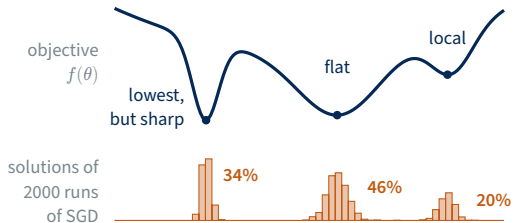
The **traditional approach** of ***"optimizing the average cost"***
is an obstacle to understanding the **robustness** of decisions and models learned from data

II ***Complex problems*** call for ***stochastic methods***, whose randomness is overlooked

Highly nonconvex landscapes



Variety of possible solutions



runs from random initializations return different solutions
these solutions may differ widely in expected performance

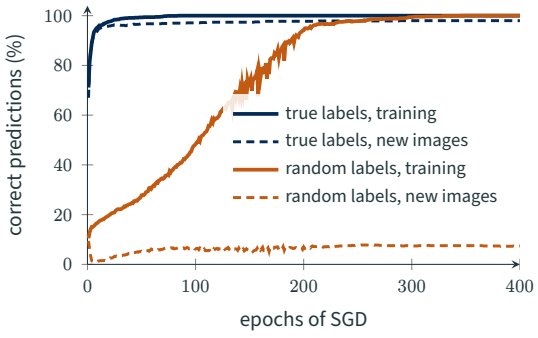
Overparametrization example: fitting random labels

- With the true labels or with labels **drawn at random**, SGD on the empirical risk can reach **zero training error** in neural networks

- **Zero training error** on random labels suggests that the model has VC-dim $\geq n$...
- ... and classical learning theory bounds the **expected excess risk** by $O\left(\sqrt{\frac{\text{VC-dim} \times \log(n)}{n}}\right)$
- Yet **in the example** on 797 new images: 98% correct for true labels, 7% for random ones

> The same network and SGD fit any labels: minimizing the empirical risk **does not explain the success** on new images nor does capacity-based theory

A network with $d \approx 3 \cdot 10^5$ parameters, trained by SGD on $n = 1000$ images of digits (8×8 pixels, 10 classes)

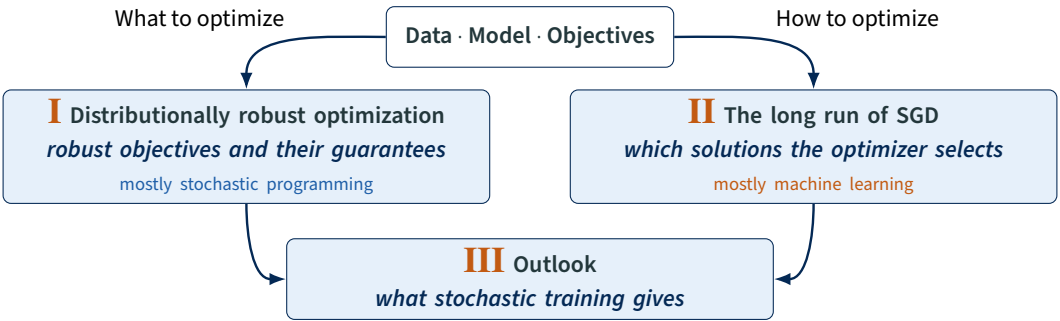


digits dataset of `scikit-learn`, see [Zhang et al.'17] on CIFAR-10

[Zhang et al.'17]

This talk

Challenge the classical *"optimizing the average cost"* **paradigm**



Joint work with W. Azizian, J. Malick and P. Mertikopoulos;
with C. Dapogny, A. Meda, J. Prando and B. Thibert (shape optimization);
with F. Vincent (`skwdro`), P.-L. Cauvin (games);
and with P. Escande and I. Zekhnini (minimizers of empirical risks)

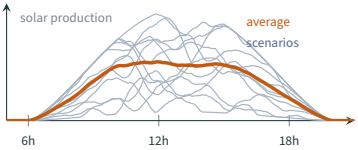
I

Distributionally robust optimization

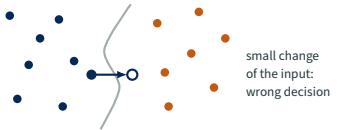
What to optimize: robust objectives, their guarantees and their computation

Why robustness?

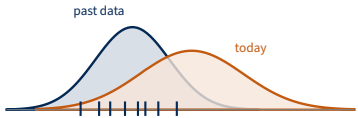
Difficult-to-predict environments



Attacks against complex models



Biased, outdated, insufficient data



Regulations

Power systems: $N-1$ security, the network must withstand the loss of any single element

Artificial intelligence: the EU AI Act requires appropriate accuracy, robustness and cybersecurity for high-risk systems

Question What to optimize to obtain decisions and models that **remain good** on new data?

[Ben-Tal, El Ghaoui & Nemirovski'09] [Kolter & Madry'18] [EU AI Act, Art. 15]

Optimization under uncertainty

- Decision $\theta \in \Theta$ has cost $\ell(\theta, \xi)$ under the **uncertain variable** $\xi \in \Xi$
- Samples $\xi_1, \dots, \xi_n \overset{\text{iid}}{\sim} P$, empirical distribution $\hat{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$
- ***Robust optimization*** (RO): the **worst cost** over a set of values of ξ
- ***Distributionally robust optimization*** (DRO): the **worst average cost** over a set of distributions

Sample average (SAA)

$$\mathbb{E}_{\xi\sim\widehat{P}_n}[\ell(\theta,\xi)]$$



every sample has weight $1/n$: optimistic (the optimizer's curse)

Robust optimization

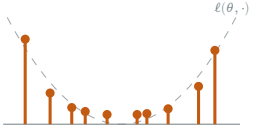
$$\sup_{\xi\in U} \ell(\theta,\xi)$$



all the weight on the **worst point**: pessimistic, U is hard to design

Kullback–Leibler DRO

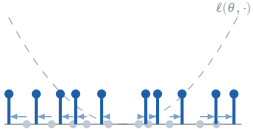
$$\sup_{Q:\,\mathrm{KL}(Q\|\widehat{P}_n)\le\rho} \mathbb{E}_{\xi\sim Q}[\ell(\theta,\xi)]$$



reweights the samples: more weight on the high costs

Wasserstein DRO

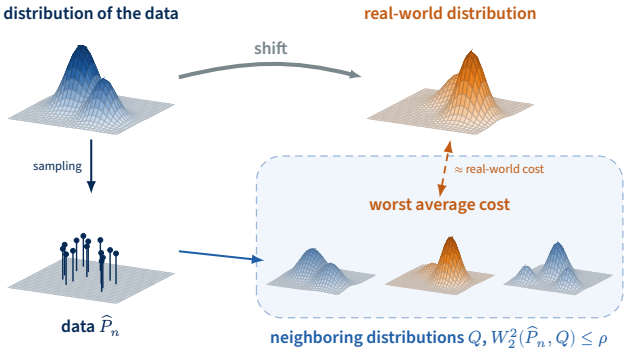
$$\sup_{Q:\,W_2^2(\widehat{P}_n,Q)\le\rho} \mathbb{E}_{\xi\sim Q}[\ell(\theta,\xi)]$$



moves the samples to higher costs, possibly to new points

[Scarf'58] [Ben-Tal, El Ghaoui & Nemirovski'09] [Delage & Ye'10] [Ben-Tal et al.'13] [Esfahani & Kuhn'18] [Kuhn, Shafiee & Wiesemann'25]

Wasserstein distributionally robust optimization



- The data define a **set of possible real-world distributions**
- The **worst average cost** over this set ≈ the real-world cost

Wasserstein DRO (WDRO)

$$\min_{\theta} \sup_{Q:\, W_2^2(\widehat{P}_n, Q) \le \rho} \mathbb{E}_{\xi \sim Q}\left[\ell(\theta, \xi)\right]$$

instead of the sample average $\min_\theta \widehat{R}_n(\theta)$

- Optimal transport cost $W_2^2(\hat{P}_n, Q) := \min\{\mathbb{E}_{(\xi,\zeta)\sim\pi}\|\xi - \zeta\|^2 : \pi \text{ with marginals } \hat{P}_n \text{ and } Q\}$: the neighborhood contains distributions supported on ***new points***
- ***Generalization guarantees*** and ***resilience to shifts***

[Esfahani & Kuhn'18] [Blanchet & Murthy'19] [Gao & Kleywegt'23] **[Azizian, I. & Malick'23a,b]**

Mathematical aspects: a linear program

- For a fixed θ, the inner problem defines the **robust risk**

$$\begin{aligned}\widehat{\mathcal{R}}_\rho(\theta) &:= \sup_{Q\in\mathcal{P}(\Xi):\ W_2^2(\widehat{P}_n,Q)\le\rho} \mathbb{E}_{\xi\sim Q}\big[\ell(\theta,\xi)\big] \\ &= \sup_{\pi\in\mathcal{P}(\Xi\times\Xi):\ \pi_1=\widehat{P}_n,\ \mathbb{E}_\pi\|\xi-\zeta\|^2\le\rho} \mathbb{E}_{(\xi,\zeta)\sim\pi}\big[\ell(\theta,\zeta)\big] \qquad \text{the distribution } Q=\pi_2 \text{ disappears} \\ &= \sup_{\pi\in\mathcal{M}_+(\Xi\times\Xi):\ \pi_1=\widehat{P}_n,\ \langle\pi,c\rangle\le\rho} \langle\pi,\tilde{\ell}_\theta\rangle\end{aligned}$$

with the pairing $\langle\pi,\varphi\rangle := \int\varphi\,\mathrm{d}\pi$ between measures and continuous functions on a compact Ξ, transport cost $c(\xi,\zeta) := \|\xi-\zeta\|^2$, and $\tilde{\ell}_\theta(\xi,\zeta) := \ell(\theta,\zeta)$

- This is a **linear program** in the measure π, with n mass constraints and one budget constraint: the ***worst-case distribution*** $\pi_2^\star$ is supported on at most $n+1$ points

[Pinelis'16] [Yue, Kuhn & Wiesemann'22]

Computational aspects: duality

- Lagrangian duality on the budget constraint, then the sup over each conditional measure $\pi(\cdot|\xi_i)$ of the transport plan is attained at a Dirac mass: for a transport cost c,

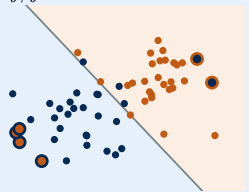
$$\widehat{\mathcal{R}}_\rho(\theta) = \inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^{n} \sup_{\zeta \in \Xi} \left[\ell(\theta, \zeta) - \lambda\, c(\xi_i, \zeta) \right] \right\}$$

- Solvable when the inner sup is tractable
  - ***Tractable***: piecewise-linear costs and linear models (newsvendor, least squares, logistic regression), as disciplined convex programs
  - ***Intractable***: complex costs and models (nonconvex inner sup over ζ)
- The cost c is adapted to the data and to the growth of the loss

Example: logistic regression labels $y_i \in \{-1, +1\}$

cost $c((x, y), (x', y')) = \|x - x'\| + \kappa\ \mathbb{1}_{y \neq y'}$ (κ: price of a flip)

$$\begin{aligned} \min_{\theta, \lambda, s}\ & \lambda\rho + \frac{1}{n}\sum_{i=1}^{n} s_i \\ \text{s.t.}\ & \log\left(1 + e^{-y_i \langle x_i, \theta \rangle}\right) \leq s_i \\ & \log\left(1 + e^{y_i \langle x_i, \theta \rangle}\right) - \kappa\lambda \leq s_i \\ & \|\theta\|_* \leq \lambda \end{aligned}$$



worst-case distribution ($n = 60$): labels flipped (circled) far from the boundary

[Shafieezadeh-Abadeh, Esfahani & Kuhn'15] [Esfahani & Kuhn'18] [Blanchet & Murthy'19] [Gao & Kleywegt'23]

Entropic regularization: a differentiable approximation

- The transport plan of the problem is penalized by its **relative entropy** with respect to a reference $\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) = \hat{P}_n(\mathrm{d}\xi)\,\pi_0(\mathrm{d}\zeta|\xi)$, where $\pi_0(\mathrm{d}\zeta|\xi) \propto e^{-\|\xi-\zeta\|^2/(2\sigma^2)}\,\mathbb{1}_\Xi(\zeta)\,\mathrm{d}\zeta$

$$\widehat{\mathcal{R}}_\rho^\varepsilon(\theta) := \sup_{\pi \in \mathcal{M}_+(\Xi\times\Xi):\ \pi_1 = \widehat{P}_n,\ \langle \pi, c\rangle \le \rho} \langle \pi, \tilde{\ell}_\theta\rangle - \varepsilon\ \mathsf{KL}(\pi \| \pi_0)$$

$$= \inf_{\lambda \ge 0}\ \lambda\rho + \frac{\varepsilon}{n}\sum_{i=1}^n \log \mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi_i)}\left[e^{\left(\ell(\theta,\zeta)-\lambda\|\xi_i-\zeta\|^2\right)/\varepsilon}\right]$$

- As in the ***Sinkhorn*** distances of optimal transport, but with only one fixed marginal: the sup over ζ becomes a soft-max over samples of π_0, ***smooth*** in (θ, λ)

Theorem [Azizian, I. & Malick'23a]

If $\Xi \subset \mathbb{R}^d$ is compact, convex, with nonempty interior, and $\ell(\theta, \cdot)$ is Lipschitz, then as $\varepsilon \to 0$,

$$0 \le \widehat{\mathcal{R}}_\rho(\theta) - \widehat{\mathcal{R}}_\rho^\varepsilon(\theta) \le O\big(\varepsilon\, d \log(1/\varepsilon)\big)$$

[Cuturi'13] **[Azizian, I. & Malick'23a]** with the entropy in the constraint (Sinkhorn DRO): [Wang, Gao & Xie'25]

Statistical aspects: generalization

- **Direct approach**: concentration of the empirical distribution, $\mathbb{P}[W_2^2(\hat{P}_n, P) \leq \rho] \geq 1 - c_1 e^{-c_2\, n\, \rho^{d/2}}$
  - With $d > 4$ and $\rho \propto \left(\log(1/\delta)/n\right)^{2/d}$, with probability at least $1-\delta$, for all θ, $\quad R(\theta) \leq \widehat{\mathcal{R}}_\rho(\theta)$
  - Overly pessimistic: the number of samples grows exponentially with the dimension d

Theorem [Azizian, I. & Malick'23b]

Under regularity assumptions on the model class, there are $\rho_n = O(\sqrt{(1+\log(1/\delta))/n})$ and a critical radius $\rho_c > 0$ such that, for $\rho_n \leq \rho \leq \rho_c/2$, with probability at least $1-\delta$, for all θ,

$$\mathbb{E}_{\xi\sim Q}[\ell(\theta,\xi)] \leq \widehat{\mathcal{R}}_\rho(\theta) \qquad \text{for } Q = P \text{ and for all } Q \text{ close enough to } P$$

Similar generalization guarantees hold for the regularized version with $\varepsilon, \sigma \propto \rho$

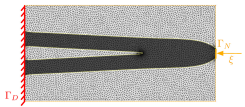
A radius of order $1/\sqrt{n}$ gives an **exact upper bound** on the true risk

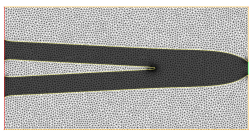
[Fournier & Guillin'15] [Esfahani & Kuhn'18] **[Azizian, I. & Malick'23b]**

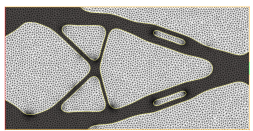
DRO example: shape optimization

Uncertain loads

Cantilever beam: minimize the compliance under a volume constraint, when the load ξ applied on Γ_N is uncertain around $(-1, 0)$





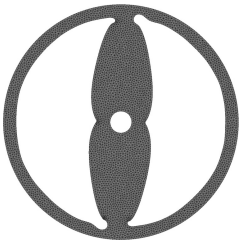


$\rho = 0$ $\rho = 2$

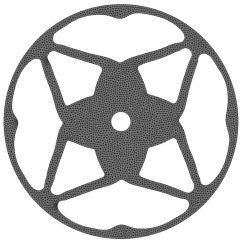
As ρ grows, cross members appear, which resist loads in other directions

Uncertain geometry

Rotating wheel: the manufactured shape is the blueprint Ω moved along its normal by a random field with 10 random coefficients



optimal for the blueprint



distributionally robust

To first order in the perturbation, the robust objective has a closed form

[Dapogny et al.'23] [Dapogny et al.'26]

WDRO in practice: the skwdro library

- The entropic approximation turns WDRO into a **new loss**, trainable by gradient methods
- Open-source library ***skwdro***: `https://github.com/iutzeler/skwdro`
  - Interfaces for `scikit-learn` estimators, as `dro` [Wang et al.'25], and for `PyTorch` models

robust `scikit-learn` estimators

```
- from sklearn.linear_model import LinearRegression
+ from skwdro.linear_models import LinearRegression

  # Fit the model
- model = LinearRegression()
+ model = LinearRegression(rho=0.1)  # radius
  model.fit(X_train, y_train)

  # Predict the target values
  y_pred = model.predict(X_test)
```

robust PyTorch losses

```
+ rho = pt.tensor(.1)  # Wasserstein radius
+ robust_loss = robustify(loss_fn, model, rho,
+                         x_init, y_init)

  for batch_x, batch_y in train_loader:
      optimizer.zero_grad()
-     loss = loss_fn(model(batch_x), batch_y).mean()
+     loss = robust_loss(batch_x, batch_y)
      loss.backward()
      optimizer.step()
```

compared with the code without robustness: lines removed (–) and added (+)

[Vincent et al.'26] [Wang et al.'25]

Robust objectives for neural networks

- To first order in $\sqrt{\rho}$, WDRO is a **gradient penalty**

$$\widehat{\mathcal{R}}_\rho(\theta) = \widehat{R}_n(\theta) + \sqrt{\rho}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_\xi \ell(\theta, \xi_i)\right\|^2\right)^{1/2} + o(\sqrt{\rho})$$

> WDRO penalizes the **sensitivity of the loss to moving the samples**

- For neural networks, robust objectives are **difficult to optimize**
  - Computational cost $\approx$ that of ERM $\times$ number of samples of π_0, and poor conditioning ($\varepsilon \propto \rho$)
  - Nonconvex problems: the solution returned depends on the optimizer

In nonconvex problems, **the optimizer selects the solution**.
Which solutions does stochastic gradient descent select?

[Bartl et al.'21] [Gao, Chen & Kleywegt'24] [Hu et al.'18] [Sagawa et al.'20] [Smith et al.'21] **[Azizian, I. & Malick'23a]**

II

The long run of stochastic gradient descent

How to optimize: which solutions does SGD select, and how often?

Stochastic gradient descent with constant step-size

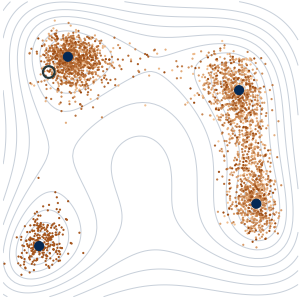
- **Objective** $f\colon \mathbb{R}^d \to \mathbb{R}$, smooth and ***nonconvex***, e.g., the empirical risk $f = \widehat{R}_n$
- **Stochastic gradient descent** (SGD) with **constant step-size** $\eta > 0$

$$\theta_{k+1} = \theta_k - \underbrace{\eta}_{\text{step-size}} \Big[\nabla f(\theta_k) + \underbrace{Z(\theta_k; \omega_{k+1})}_{\text{zero-mean noise}}\Big]$$

For $f = \widehat{R}_n$ and minibatches ω of B samples,
$Z(\theta;\omega) = \frac{1}{B}\sum_{i\in\omega} \nabla_\theta \ell(\theta, \xi_i) - \nabla \widehat{R}_n(\theta)$

- ***No pointwise convergence***, since the exploration due to the noise does not vanish, but ***very efficient*** in practice

Running example: Himmelblau function $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$, with four global minimizers (dots)



one run of SGD (Gaussian noise), 600 000 steps, from light to dark (start circled)

Question In the long run, which minimizers does SGD visit, and how often?

Why large deviations?

- Classical guarantees describe SGD **on average** (convergence to critical points, avoidance of saddle points), **not which minimizer it selects**
- **Two scales of noise**, over the time $t = \eta k$ of the gradient flow $\dot{\theta} = -\nabla f(\theta)$

Langevin (sampling)	$\theta_{k+1} = \theta_k - \eta \nabla f(\theta_k) + \sqrt{2\eta}\,\sigma\,\varepsilon_{k+1}$	accumulated noise of order $\sigma\sqrt{t}$
SGD	$\theta_{k+1} = \theta_k - \eta \nabla f(\theta_k) - \eta\, Z(\theta_k; \omega_{k+1})$	accumulated noise of order $\sigma\sqrt{\eta t}$

> As $\eta \to 0$, SGD is a **small random perturbation** of the gradient flow: leaving the basin of a minimizer is a ***rare event***, with a probability **exponentially small** in $1/\eta$

- Stochastic differential equations approximate SGD on **finite horizons**, not on the **exponentially long times** that decide the long run
- Motivates the study of the **invariant measure** μ_η of SGD: $\quad \theta \sim \mu_\eta \implies \theta - \eta\big[\nabla f(\theta) + Z(\theta;\omega)\big] \sim \mu_\eta$

[Ghadimi & Lan'13] [Brandière & Duflo'96] [Mertikopoulos et al.'20] [Freidlin & Wentzell'12] [Li, Tai & E'17,'19] [Douc et al.'18]

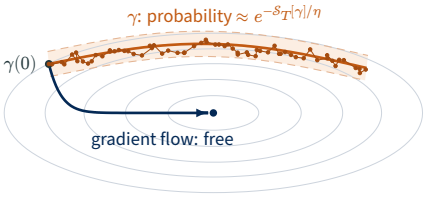
A large deviation principle for SGD

- **Ingredients** for comparing SGD with a smooth curve $\gamma\colon [0,T] \to \mathbb{R}^d$
  - ***Hamiltonian***, the cumulant generating function of the stochastic gradient $\mathcal{H}(\theta, p) := \log \mathbb{E}\big[\exp\langle p, \nabla f(\theta) + Z(\theta;\omega)\rangle\big]$
  - ***Lagrangian*** $\mathcal{L}(\theta, v) := \mathcal{H}^*(\theta, -v)$, its convex conjugate in p
  - ***Action*** $\mathcal{S}_T[\gamma] := \int_0^T \mathcal{L}(\gamma(t), \dot\gamma(t))\,\mathrm{d}t$

Lemma

As $\eta \to 0$, $\qquad \mathbb{P}(\text{the first } T/\eta \text{ iterates of SGD stay close to } \gamma) \approx \exp\big(-\mathcal{S}_T[\gamma]/\eta\big)$

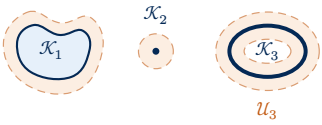
- $\mathcal{L}(\theta, v) \geq 0$, with equality if and only if $v = -\nabla f(\theta)$: ***following the gradient flow is free***, and deviating from it has an exponential cost
- ***Gaussian case*** $Z \sim \mathcal{N}(0, \sigma^2 I)$: $\quad \mathcal{L}(\theta, v) = \dfrac{\|v + \nabla f(\theta)\|^2}{2\sigma^2}$



[Kifer'88] [Freidlin & Wentzell'12] **[Azizian et al.'24]**

Transitions between critical components

- **Ingredients** for following SGD between the critical points of f
  - ***Components*** $\mathcal{K}_1, \dots, \mathcal{K}_N$ of the critical set $\mathrm{crit}(f) = \{\theta : \nabla f(\theta) = 0\}$, finitely many for definable functions
  - ***Transition cost***, or quasi-potential, $B_{ij} := \inf\{\mathcal{S}_T[\gamma] : \gamma(0) \in \mathcal{K}_i,\ \gamma(T) \in \mathcal{K}_j,\ T \in \mathbb{N}\}$
  - ***Energy*** $E_i := \min_{\mathcal{T}} \sum_{(j \to k) \in \mathcal{T}} B_{jk}$, over the spanning trees $\mathcal{T}$ rooted at i

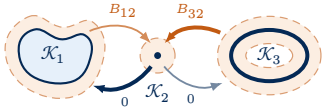


a plateau, a saddle point and a ring of minimizers, with small neighborhoods

Lemma

As $\eta \to 0$, $\mathbb{P}(\text{SGD goes from } \mathcal{K}_i \text{ to } \mathcal{K}_j) \approx \exp(-B_{ij}/\eta)$, and the chain of these transitions has the invariant distribution $\pi(i) \propto \exp(-E_i/\eta)$

- $B_{ij} = 0$ if a gradient flow goes from $\mathcal{K}_i$ to $\mathcal{K}_j$: ***going down is free***, and climbing has an exponential cost
- E_i: smallest total cost of transitions leading all the other components to $\mathcal{K}_i$



thick: a tree of minimal cost rooted at $\mathcal{K}_1$, so that $E_1 = B_{32}$

[Kifer'88] [Freidlin & Wentzell'12] **[Azizian et al.'24]**

Long-run distribution of SGD

Theorem [Azizian et al.'24]

Let $\varepsilon > 0$ and $\mathcal{U}_1, \dots, \mathcal{U}_N$ be small enough neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_N$. For $\eta > 0$ small enough, with constants $c, c_0 > 0$,

- **Concentration on the critical set**: $\mu_\eta(\mathcal{U}_1 \cup \cdots \cup \mathcal{U}_N) \geq 1 - e^{-c/\eta}$
- **Boltzmann–Gibbs distribution**: $\mu_\eta(\mathcal{U}_i) \propto \exp\big(-(E_i + O(\varepsilon))/\eta\big)$ for all i
- **Concentration on ground states**: $\mu_\eta(\mathcal{U}_0) \geq 1 - e^{-c_0/\eta}$, for $\mathcal{U}_0$ a neighborhood of $\bigcup_{i \in \arg\min E} \mathcal{K}_i$

- **Assumptions**: f smooth and coercive, with a coercive gradient; noise with zero mean, non-degenerate covariance and sub-Gaussian tails
- **Global convergence time**: the same transition costs give sharp estimates of the time to reach a global minimizer

> In the long run, SGD visits each critical component with a probability set by an **energy that depends on the loss and on the noise**

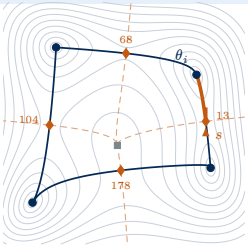
[Azizian et al.'24] [Azizian et al.'25]

Isotropic Gaussian noise

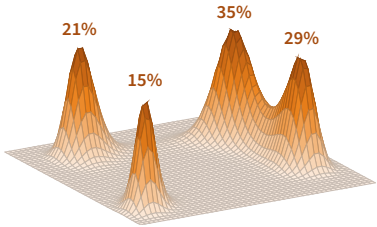
- Isotropic Gaussian noise $Z \sim \mathcal{N}(0, \sigma^2 I)$: climbing from a minimizer θ_i to a neighboring saddle point s costs $2\big(f(s) - f(\theta_i)\big)/\sigma^2$, going down is free, and

$$E_i = \frac{2f(\theta_i)}{\sigma^2} + \text{const}, \qquad \text{so that} \qquad \mu_\eta(\mathcal{U}_i) \propto \exp\Big(-\frac{2f(\theta_i)}{\eta\sigma^2}\Big)$$

A Gibbs measure with temperature $\eta\sigma^2/2$: SGD concentrates on the **global minimizers**



minimizers (navy), saddle points (orange, with $f(s)$), free descents (navy), basin boundaries (dashed), a climb from θ_i to s (orange)

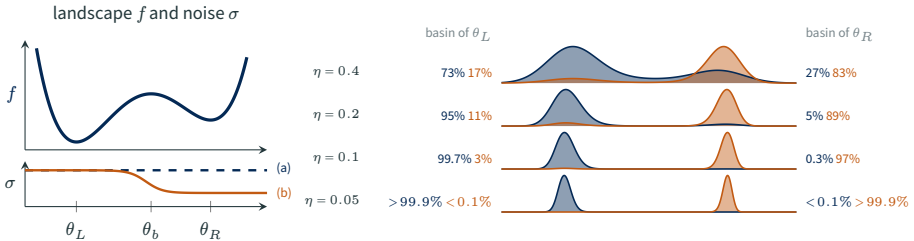


long-run distribution of SGD ($\eta = 0.001$, $\eta\sigma^2/2 = 15$, 25 000 000 iterates): the shares of the basins are those of the Gibbs measure, up to 1%

What happens when the variance changes across the landscape?

- Landscape f on $\mathbb{R}$ with two minimizers, θ_L (global) and θ_R, and Gaussian noise of standard deviation $\sigma(\theta)$: leaving the basin of θ_m over the barrier θ_b costs

$$B_m = 2\left|\int_{\theta_m}^{\theta_b} \frac{f'(\theta)}{\sigma(\theta)^2}\,\mathrm{d}\theta\right|, \qquad \text{and the energies are} \quad E_L = B_R, \quad E_R = B_L$$



(a) constant noise: $E_L = 0.73 < E_R = 1.34$ (b) less noise around θ_R: $E_L = 1.71 > E_R = 1.41$

Rows: long-run distribution of SGD for decreasing η, with the probability of each basin

With less noise around θ_R, SGD selects θ_R, although $f(\theta_R) > f(\theta_L)$

[Azizian et al.'24] for noise that grows with the loss, see also [Mori et al.'22]

Implicit regularization: GD, SGD and Wasserstein DRO

- Empirical risk $\widehat{R}_n$, per-sample gradients $\nabla_\theta \ell(\theta, \xi_i)$ with empirical covariance Cov_n, step-size η, minibatches of $B \ll n$ samples, radius ρ

GD

follows the gradient flow of

$$\widehat{R}_n + \frac{\eta}{4}\|\nabla \widehat{R}_n\|^2$$

penalizes the ***mean gradient***, zero at the critical points of $\widehat{R}_n$

SGD

follows on average the flow of

$$\widehat{R}_n + \frac{\eta}{4}\|\nabla \widehat{R}_n\|^2 + \frac{\eta}{4B}\,\mathrm{tr}\,\mathrm{Cov}_n$$

also penalizes the ***disagreement between per-sample gradients***, zero only if every per-sample gradient vanishes

Wasserstein DRO

minimizes, to first order in $\sqrt{\rho}$,

$$\widehat{R}_n + \sqrt{\rho}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla_x \ell(\theta, \xi_i)\|^2\right)^{1/2}$$

penalizes the ***sensitivity to moving the features*** x of the samples $\xi = (x, y)$

GD and SGD penalize gradients in the **parameters** θ, Wasserstein DRO gradients in the **data** x

[Barrett & Dherin'21] [Smith et al.'21] [Bartl et al.'21] [Gao, Chen & Kleywegt'24] precise statements in backup

SGD controls the Wasserstein penalty

- Let the loss depend on the features **only through a fully connected layer** $u = W_1 x + b_1$: $\ell(\theta, (x, y)) = L(W_1 x + b_1, y)$, with L given by the other layers, so that $\nabla_x \ell = W_1^\top \nabla_u L = W_1^\top \nabla_{b_1} \ell$. As $\nabla_{b_1} \ell$ is a part of the per-sample gradient $\nabla_\theta \ell$, $\quad \|\nabla_x \ell\| \le \|W_1\|_{\mathsf{op}} \|\nabla_{b_1} \ell\| \le \|W_1\|_{\mathsf{op}} \|\nabla_\theta \ell\|$
- Then, by the variance identity, for all θ,

$$\frac{1}{n} \sum_{i=1}^n \|\nabla_x \ell(\theta, \xi_i)\|^2 \;\le\; \|W_1\|_{\mathsf{op}}^2 \, \frac{1}{n} \sum_{i=1}^n \|\nabla_\theta \ell(\theta, \xi_i)\|^2 \;=\; \|W_1\|_{\mathsf{op}}^2 \Big(\|\nabla \widehat{R}_n(\theta)\|^2 + \mathsf{tr}\, \mathsf{Cov}_n \left[\nabla_\theta \ell(\theta, \xi)\right] \Big)$$

- With $\sqrt{ab} \le (a+b)/2$, for all θ and all radii $\rho > 0$,

$$\underbrace{\widehat{R}_n + \sqrt{\rho} \left(\frac{1}{n} \sum_{i=1}^n \|\nabla_x \ell(\theta, \xi_i)\|^2 \right)^{1/2}}_{\text{Wasserstein DRO, first order}} \le \underbrace{\widehat{R}_n + \frac{\eta}{4} \|\nabla \widehat{R}_n\|^2 + \frac{\eta}{4B} \, \mathsf{tr}\, \mathsf{Cov}_n \left[\nabla_\theta \ell\right]}_{\text{modified loss of SGD}} + \frac{\rho B}{\eta} \|W_1\|_{\mathsf{op}}^2$$

Unlike GD, SGD minimizes an upper bound on the first-order Wasserstein objective, up to a term in $\|W_1\|_{\mathsf{op}}^2$: its noise scale η/B plays the role of the radius

[Smith et al.'21] [Bartl et al.'21] [Gao, Chen & Kleywegt'24]

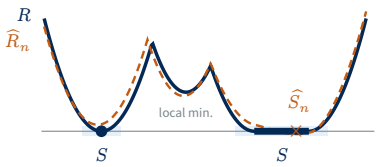
III

Outlook

What does stochastic training give?

When $n \gg d$, nonconvex empirical risk minimizers concentrate

- Parameters $\theta \in \mathbb{R}^d$, minimizers $S := \arg\min R$ of the true risk (possibly unbounded, with several components), and $\hat{S}_n := \arg\min (\widehat{R}_n + \lambda_n \|\cdot\|^2)$, $\lambda_n \propto \sqrt{d/n}$
- **Assumptions**: sub-Gaussian increments of the loss in θ, quadratic growth of R around S



the empirical minimizers $\widehat{S}_n$ lie in thin tubes around S

Theorem [Escande, I. & Zekhnini'26]

With high probability, $\hat{S}_n$ stays in a ball that does not depend on n, and, if S is the union of m smooth components of dimension at most d_S,

$$\sup_{\theta \in \widehat{S}_n} \operatorname{dist}(\theta, S)^2 \lesssim \frac{d + d_S \log n + \log m}{n}$$

Every global minimizer of the regularized empirical risk is close to a true minimizer when $n \gg d$

[Rockafellar & Wets'98] [Mei, Bai & Montanari'18] [Shapiro, Dentcheva & Ruszczyński'21] **[Escande, I. & Zekhnini'26]**

When $d \gg n$, exact empirical risk minimizers may be useless

Question When $d \gg n$, exact minimizers can be useless (a diffusion model would reproduce its training images), but **stochastic training does not compute them**. What does it optimize?

- A minibatch drawn with replacement **reweights the data at random**

$$\frac{1}{B}\sum_{b=1}^{B} \nabla_\theta \ell(\theta, \xi_{i_b}) = \sum_{i=1}^{n} q_i \, \nabla_\theta \ell(\theta, \xi_i), \qquad q = \frac{(c_1, \dots, c_n)}{B}, \quad c \sim \text{Multinomial}\big(B; \tfrac{1}{n}, \dots, \tfrac{1}{n}\big)$$

so that one epoch of minibatches is a ***bootstrap resample*** of the data



SGD takes gradient steps on **randomly reweighted empirical risks**, whereas Kullback–Leibler DRO (Part I) takes the worst reweighting in a ball

[Efron'79] [Kamb & Ganguli'25] [Bonnaire et al.'25] [Lan, Liu & Schäfer'25]

Towards a statistical theory of stochastic training

- The long-run distribution of SGD is governed by **energies that depend on the loss and on the noise**. These energies have no explicit statistical meaning yet
- The noise of SGD comes from **random reweightings** of the data, so that its rare events are **unusual reweightings**

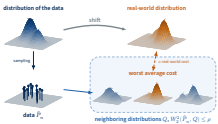
> The **long run of SGD** meets the **statistics of resampling**: for instance, SGD tends to forget the structures that rely on a few samples

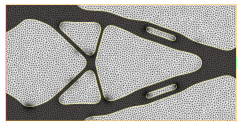
Statistics	Time	Consequences
Which solutions survive the resampling of the data, and for how long?	What does training forget, and when can it stop without validation data?	Generalization, small groups in the data, memorization in generative models

Conclusion

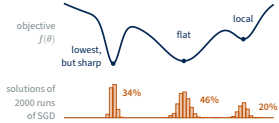
The **traditional approach** of ***"optimizing the average cost"***
is an obstacle to understanding the **robustness** of decisions and models learned from data

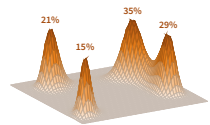
- **Robustness by design**: Wasserstein distributionally robust optimization
  - Exact generalization guarantees with a radius of order $1/\sqrt{n}$
  - Tractable through duality or entropic regularization, open-source library `skwdro`
  - Effective for linear models and engineering design, costly, sometimes disappointing for neural networks
- **Robustness through the optimizer**: stochastic gradient descent
  - Long-run distribution $\propto \exp(-E_i/\eta)$, with energies that depend on the loss and on the noise
  - Low-noise solutions are favored, and SGD controls the first-order Wasserstein penalty
- **Outlook**: the statistics of data-driven optimization, from the minimizers of SAA and ERM problems to the trajectories of SGD











Thank you for your attention

www.iutzeler.org

Distributional robustness

Azizian, I., Malick. ***Regularization for Wasserstein distributionally robust optimization***. ESAIM: COCV, 2023.

Azizian, I., Malick. ***Exact generalization guarantees for (regularized) Wasserstein distributionally robust models***. NeurIPS, 2023.

Dapogny, I., Meda, Thibert. ***Entropy-regularized Wasserstein distributionally robust shape and topology optimization***. Struct. Multidiscip. Optim., 2023.

Dapogny, I., Prando, Thibert. ***Distributionally robust shape and topology optimization: overview and deployment for geometric uncertainties***. Preprint, 2026.

Vincent, Azizian, I., Malick. ***skwdro: a library for Wasserstein distributionally robust machine learning***. JMLR (MLOSS), 2026.

Stochastic gradient descent and statistics of learning

Azizian, I., Malick, Mertikopoulos. ***What is the long-run distribution of stochastic gradient descent? A large deviations analysis***. ICML, 2024.

Azizian, I., Malick, Mertikopoulos. ***The global convergence time of stochastic gradient descent in non-convex landscapes: sharp estimates via large deviations***. ICML, 2025.

Azizian, Cauvin, I., Malick, Mertikopoulos. ***The long-run distribution of regularized learning in non-concave games: a large deviations approach***. NeurIPS, 2026.

Escande, I., Zekhnini. ***Finite-sample concentration of the minimizers of nonconvex empirical risks***. Preprint, 2026.

Backup slides

Assumptions for the large deviations of SGD

- **Objective function f**
  - ***smooth***: C^2, and ∇f is Lipschitz continuous
  - ***coercive***: $\lim_{\|\theta\|\to\infty} f(\theta) = +\infty$, ***gradient coercive***: $\lim_{\|\theta\|\to\infty} \|\nabla f(\theta)\| = +\infty$
  - ***critical set***: finitely many smoothly connected components $\mathcal{K}_1, \dots, \mathcal{K}_N$
- **Noise Z**
  - ***proper***: $\mathbb{E}[Z(\theta;\omega)] = 0$ and $\mathrm{Cov}[Z(\theta;\omega)] \succ 0$ for all $\theta \in \mathbb{R}^d$
  - ***limited growth***: Z is C^2 and $Z(\theta;\omega) = O(\|\theta\|)$ almost surely
  - ***sub-Gaussian***: $\log \mathbb{E}\big[\exp\langle p, Z(\theta;\omega)\rangle\big] \le \sigma^2\|p\|^2/2$

Example Regularized ERM, $f(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta,\xi_i) + \frac{\lambda}{2}\|\theta\|^2$, with SGD sampling one example uniformly at random: $Z(\theta;\omega) = \nabla_\theta \ell(\theta,\xi_\omega) - \frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell(\theta,\xi_i)$. Its covariance is non-degenerate only if the centered per-sample gradients span $\mathbb{R}^d$, which requires $d < n$.

Proof sketch: from discrete to continuous time

- **Discrete time** $\theta_{k+1} = \theta_k - \eta[\nabla f(\theta_k) + Z(\theta_k; \omega_{k+1})]$
- **Continuous time**
  - ***interpolated trajectory***: for $t \in [\eta k, \eta(k+1)]$, $\quad X_t = \theta_k + (t/\eta - k)(\theta_{k+1} - \theta_k)$
  - ***continuous trajectory with discretized noise***, with $Y_0 = \theta_0$, $\quad \dot{Y}_t = -\nabla f(Y_t) - Z(Y_t; \omega_{\lfloor t/\eta \rfloor + 1})$
  - the two trajectories stay at distance $O(\eta)$ on $[0, T]$
- **Large deviations**, following Freidlin and Wentzell: the trajectory Y is a point in the space of continuous curves $\mathcal{C}([0,T], \mathbb{R}^d)$, and it satisfies a large deviation principle with rate function $\mathcal{S}_T$

Remarks Time is rescaled, $\Delta t = 1$ corresponding to $1/\eta$ iterations, to have enough noise between t and $t+1$. The SDE $\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{\eta\ \mathrm{Cov}[Z(X_t;\cdot)]}\,\mathrm{d}W_t$ approximates SGD on finite horizons, but its long-run behavior is a different question.

[Freidlin & Wentzell'12] [Kifer'88] [Li, Tai & E'17,'19] **[Azizian et al.'24]**

Implicit regularization and Wasserstein DRO: the precise statements

- **GD** [Barrett & Dherin'21] with step η follows the flow of $\widetilde{R}_{\mathsf{GD}} = \widehat{R}_n + \frac{\eta}{4}\|\nabla \widehat{R}_n\|^2$, error $O(\eta^3)$ per step
- **SGD** [Smith et al.'21]: an epoch visits $m = n/B$ minibatches, of empirical risks $\widehat{R}_1, \dots, \widehat{R}_m$ with mean $\widehat{R}_n$; on average over their order, an epoch follows, up to higher order in $m\eta$, the flow of

$$\widetilde{R}_{\mathsf{SGD}} = \widehat{R}_n + \frac{\eta}{4m}\sum_{k=1}^{m}\|\nabla \widehat{R}_k\|^2 = \widehat{R}_n + \frac{\eta}{4}\|\nabla \widehat{R}_n\|^2 + \frac{\eta}{4m}\sum_{k=1}^{m}\|\nabla \widehat{R}_k - \nabla \widehat{R}_n\|^2$$

  and a minibatch drawn without replacement has $\mathbb{E}\|\nabla \widehat{R}_k - \nabla \widehat{R}_n\|^2 = \frac{n-B}{n-1}\,\frac{1}{B}\,\mathsf{tr}\,\mathsf{Cov}_n\left[\nabla_\theta \ell(\theta, \xi)\right]$
- **Wasserstein DRO** [Bartl et al.'21] [Gao, Chen & Kleywegt'24]: with the transport cost $\|x - x'\|^2$ on the features, if $\nabla_x \ell(\theta, \cdot)$ is Lipschitz continuous, $\widehat{\mathcal{R}}_\rho(\theta) = \widehat{R}_n(\theta) + \sqrt{\rho}\,\big(\frac{1}{n}\sum_{i=1}^{n}\|\nabla_x \ell(\theta, \xi_i)\|^2\big)^{1/2} + O(\rho)$
- **The bound**, for $B \ll n$: with $V := \|\nabla \widehat{R}_n\|^2 + \mathsf{tr}\,\mathsf{Cov}_n[\nabla_\theta \ell]$, the chain rule and $\sqrt{ab} \le (a+b)/2$ give $\sqrt{\rho}\,\|W_1\|_{\mathsf{op}}\sqrt{V} \le \frac{\eta}{4B}V + \frac{\rho B}{\eta}\|W_1\|_{\mathsf{op}}^2$, and $\frac{\eta}{4B}\|\nabla \widehat{R}_n\|^2 \le \frac{\eta}{4}\|\nabla \widehat{R}_n\|^2$

Finite steps penalize the gradient norm, minibatches the variance of the per-sample gradients, Wasserstein DRO the gradients in the data

[Barrett & Dherin'21] [Smith et al.'21] [Bartl et al.'21] [Gao, Chen & Kleywegt'24]

References I: distributional robustness

[Bartl et al.'21] D. Bartl, S. Drapeau, J. Obłój, J. Wiesel. *Sensitivity analysis of Wasserstein distributionally robust optimization problems*. Proc. R. Soc. A, 2021.

[Ben-Tal, El Ghaoui & Nemirovski'09] A. Ben-Tal, L. El Ghaoui, A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[Ben-Tal et al.'13] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, G. Rennen. *Robust solutions of optimization problems affected by uncertain probabilities*. Management Science, 2013.

[Blanchet & Murthy'19] J. Blanchet, K. Murthy. *Quantifying distributional model risk via optimal transport*. Math. Oper. Res., 2019.

[Cuturi'13] M. Cuturi. *Sinkhorn distances: lightspeed computation of optimal transport*. NeurIPS, 2013.

[Delage & Ye'10] E. Delage, Y. Ye. *Distributionally robust optimization under moment uncertainty with application to data-driven problems*. Oper. Res., 2010.

[Esfahani & Kuhn'18] P. Mohajerin Esfahani, D. Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations*. Math. Program., 2018.

[Fournier & Guillin'15] N. Fournier, A. Guillin. *On the rate of convergence in Wasserstein distance of the empirical measure*. Probab. Theory Relat. Fields, 2015.

[Gao & Kleywegt'23] R. Gao, A. Kleywegt. *Distributionally robust stochastic optimization with Wasserstein distance*. Math. Oper. Res., 2023.

[Gao, Chen & Kleywegt'24] R. Gao, X. Chen, A. Kleywegt. *Wasserstein distributionally robust optimization and variation regularization*. Oper. Res., 2024.

[Hu et al.'18] W. Hu, G. Niu, I. Sato, M. Sugiyama. *Does distributionally robust supervised learning give robust classifiers?* ICML, 2018.

[Kolter & Madry'18] Z. Kolter, A. Madry. *Adversarial robustness: theory and practice*. NeurIPS tutorial, 2018.

[Kuhn, Shafiee & Wiesemann'25] D. Kuhn, S. Shafiee, W. Wiesemann. *Distributionally robust optimization*. Acta Numerica, 2025.

References II: distributional robustness (continued)

[Mak, Morton & Wood'99] W.-K. Mak, D. P. Morton, R. K. Wood. *Monte Carlo bounding techniques for determining solution quality in stochastic programs*. Oper. Res. Lett., 1999.

[Pinelis'16] I. Pinelis. *On the extreme points of moments sets*. Math. Methods Oper. Res., 2016.

[Sagawa et al.'20] S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang. *Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization*. ICLR, 2020.

[Scarf'58] H. Scarf. *A min-max solution of an inventory problem*. Studies in the Mathematical Theory of Inventory and Production, 1958.

[Shafieezadeh-Abadeh, Esfahani & Kuhn'15] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, D. Kuhn. *Distributionally robust logistic regression*. NeurIPS, 2015.

[Shapiro, Dentcheva & Ruszczyński'21] A. Shapiro, D. Dentcheva, A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

[Smith & Winkler'06] J. E. Smith, R. L. Winkler. *The optimizer's curse: skepticism and postdecision surprise in decision analysis*. Management Science, 2006.

[Wang, Gao & Xie'25] J. Wang, R. Gao, Y. Xie. *Sinkhorn distributionally robust optimization*. Oper. Res., 2025.

[Wang et al.'25] T. Wang, J. Liu, H. Lam, H. Namkoong, J. Blanchet. *DRO: a Python library for distributionally robust optimization in machine learning*. arXiv:2505.23565, 2025.

[Yue, Kuhn & Wiesemann'22] M.-C. Yue, D. Kuhn, W. Wiesemann. *On linear optimization over Wasserstein balls*. Math. Program., 2022.

References III: SGD and statistics of learning

[Barrett & Dherin'21] D. G. T. Barrett, B. Dherin. *Implicit gradient regularization*. ICLR, 2021.

[Bonnaire et al.'25] T. Bonnaire, R. Urfin, G. Biroli, M. Mézard. *Why diffusion models don't memorize: the role of implicit dynamical regularization in training*. NeurIPS, 2025.

[Brandière & Duflo'96] O. Brandière, M. Duflo. *Les algorithmes stochastiques contournent-ils les pièges?* Ann. Inst. H. Poincaré Probab. Statist., 1996.

[Douc et al.'18] R. Douc, E. Moulines, P. Priouret, P. Soulier. *Markov Chains*. Springer, 2018.

[Efron'79] B. Efron. *Bootstrap methods: another look at the jackknife*. Ann. Statist., 1979.

[Freidlin & Wentzell'12] M. I. Freidlin, A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, 3rd ed., 2012.

[Ghadimi & Lan'13] S. Ghadimi, G. Lan. *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*. SIAM J. Optim., 2013.

[Kamb & Ganguli'25] M. Kamb, S. Ganguli. *An analytic theory of creativity in convolutional diffusion models*. ICML, 2025.

[Kifer'88] Y. Kifer. *Random Perturbations of Dynamical Systems*. Birkhäuser, 1988.

[Lan, Liu & Schäfer'25] H. Lan, Y. Liu, F. Schäfer. *A bootstrap perspective on stochastic gradient descent*. arXiv:2512.07676, 2025.

[Li, Tai & E'17,'19] Q. Li, C. Tai, W. E. *Stochastic modified equations and adaptive stochastic gradient algorithms*. ICML, 2017; *Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations*. JMLR, 2019.

[Mei, Bai & Montanari'18] S. Mei, Y. Bai, A. Montanari. *The landscape of empirical risk for nonconvex losses*. Ann. Statist., 2018.

[Mertikopoulos et al.'20] P. Mertikopoulos, N. Hallak, A. Kavis, V. Cevher. *On the almost sure convergence of stochastic gradient descent in non-convex problems*. NeurIPS, 2020.

[Mori et al.'22] T. Mori, L. Ziyin, K. Liu, M. Ueda. *Power-law escape rate of SGD*. ICML, 2022.

[Rockafellar & Wets'98] R. T. Rockafellar, R. J.-B. Wets. *Variational Analysis*. Springer, 1998.

[Smith et al.'21] S. L. Smith, B. Dherin, D. G. T. Barrett, S. De. *On the origin of implicit regularization in stochastic gradient descent*. ICLR, 2021.

[Zhang et al.'17] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals. *Understanding deep learning requires rethinking generalization*. ICLR, 2017.

References IV: our works

[Azizian, I. & Malick'23a] W. Azizian, F. Iutzeler, J. Malick. *Regularization for Wasserstein distributionally robust optimization*. ESAIM: COCV, 2023.

[Azizian, I. & Malick'23b] W. Azizian, F. Iutzeler, J. Malick. *Exact generalization guarantees for (regularized) Wasserstein distributionally robust models*. NeurIPS, 2023.

[Dapogny et al.'23] C. Dapogny, F. Iutzeler, A. Meda, B. Thibert. *Entropy-regularized Wasserstein distributionally robust shape and topology optimization*. Struct. Multidiscip. Optim., 2023.

[Dapogny et al.'26] C. Dapogny, F. Iutzeler, J. Prando, B. Thibert. *Distributionally robust shape and topology optimization: overview and deployment for geometric uncertainties*. Preprint, HAL hal-05740258, 2026.

[Vincent et al.'26] F. Vincent, W. Azizian, F. Iutzeler, J. Malick. *skwdro: a library for Wasserstein distributionally robust machine learning*. JMLR (MLOSS), 2026.

[Azizian et al.'24] W. Azizian, F. Iutzeler, J. Malick, P. Mertikopoulos. *What is the long-run distribution of stochastic gradient descent? A large deviations analysis*. ICML, 2024.

[Azizian et al.'25] W. Azizian, F. Iutzeler, J. Malick, P. Mertikopoulos. *The global convergence time of stochastic gradient descent in non-convex landscapes: sharp estimates via large deviations*. ICML, 2025.

[Azizian et al.'26] W. Azizian, P.-L. Cauvin, F. Iutzeler, J. Malick, P. Mertikopoulos. *The long-run distribution of regularized learning in non-concave games: a large deviations approach*. NeurIPS, 2026.

[Escande, I. & Zekhnini'26] P. Escande, F. Iutzeler, I. Zekhnini. *Finite-sample concentration of the minimizers of nonconvex empirical risks*. Preprint, 2026.